VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Cheminių junginių kristalografinės ir kristalocheminės informacijos išgavimas iš mokslinių straipsnių

Extracting Crystallographic Information about Chemical Compounds from Research Papers

Magistro baigiamasis darbas

Atliko:	Tomas Kovtun	(parašas)
Darbo vadovas:	prof. dr. Saulius Gražulis	(parašas)
Recenzentas:	dr. Linas Petkevičius	(parašas)

Vilnius – 2023

Santrauka

Šis darbas yra orientuotas automatizuoti kristalografinės informacijos išgavimą iš mokslinių straipsnių tekstų. Darbas apžvelgia taikytinus kalbos nuskaitymo metodus ir esamus panašių užduočių sprendimus. Taip pat yra pateikiamas naujas junginių ir parametrų siejimo metodas pagrįstas išgaunamųjų klausimų-atsakymų modeliais. Papildomai darbe yra aprašytas metodas, kuriuo buvo automatiškai anotuoti dalis apmokymo ir tikrinimo duomenų rinkinių remiantis ”Crystallography Open Database” esančiais įrašais. Gautų duomenų rinkinių pagalba buvo apmokytos trys BERT architektūros kalbos modelių realizacijos. Geriausius tikslumo rodiklius parodė konvejeris paremtas BioBERT kalbos modeliu (91,9 tikslumo ir 63,7 atkūrimo rodikliai). Nors gautas sprendimas pakankamai tiksliai išgauna su junginiu susijusius parametrus, visgi liko keli kristalografinių tekstų aspektai, kurie nėra deramai apdorojami (pvz. kristalografinių parametrų pateikimas skirtingomis slėgio ir temperatūros sąlygomis). Dalis tokių aspektų gali būti apdorojami toliau plečiant klausimų-atsakymų modelio pritaikymo sritį.

Summary

This work focuses on automating the extraction of crystallographic information from scientific articles. The work reviews applicable natural language processing methods and existing solutions for similar tasks. Additionally, the document presents a new compound-parameter association method based on extractive question answering models. Moreover, the document describes a method that automatically annotates training and validation datasets based on data from the Crystallography Open Database. Three implementations of language models of the BERT architecture were trained using the obtained datasets. The pipeline based on the BioBERT language model showed the best precision rates (91.9 precision and 63.7 recall). Although the resulting solution extracts the parameters related to the compound with sufficient precision, there are still several aspects of the crystallographic texts that are not properly addressed (e.g. presentation of the crystallographic parameters under different pressure and temperature conditions). Some of these aspects can be handled by further expansion and improvement of the question-answer model.

Turinys

Įvadas

Šis darbas yra skirtas paruošti sprendimą, gebantį papildyti kristalografinės informacijos duomenų bazę jau ištirtais junginiais ir jų kristalografinėmis savybėmis. Junginiai ir jų parametrai bus išgaunami iš laisvai prieinamų kristalografinių mokslinių straipsnių.

Kristalografijos mokslo šaka tiria kristalus, jų struktūrą, savybes ir dėsnius. Išvados daromos kristalografinių tyrimų būdu leidžia geriau suprasti medžiagas ir tolimesnes jų panaudojimo galimybes. Teoriniai skaičiavimai šioje srityje kol kas nėra pakankamai pažengę, todėl tikslią kristalografinę informaciją galima išgauti tik eksperimentiniu būdu. Tokio tipo tyrimai yra ganėtinai sudėtingi ir reikalauja tam skirtos įrangos bei kompetencijų. Kristalografiniai tyrimai turi aukštą mokslinę vertę, todėl dauguma jų būna kruopščiai dokumentuojami ir publikuojami profiliniuose moksliniuose žurnaluose. Tikslų publikacijų skaičių nustatyti yra sudėtinga, tačiau mastą leidžia suvokti bendro biomedicininių straipsnių srauto dažnis. Pavyzdžiui, 2010 metais PubMed paieškos sistema indeksavo po vieną naują biomedicininį straipsnį kas minutę (https://duncan.hull.name). Kadangi šis skaičius auga, apdoroti didėjančius informacijos srautus tampa vis sudėtingiau.

Ilgą laiką kristalografinių tyrimų straipsniai buvo publikuojami išskirtinai žmogui skirtu teksto pavidalu be griežtų formalizavimo taisyklių. Crystallographic Information File (CIF) [HAB91] formatas yra vienas iš sėkmingų, plačiai naudojamų metodų formalizuoti straipsnyje esamus duomenis kompiuteriui suprantama kalba. Besivystančios internetinės technologijos nukreipė cheminės informacijos atvaizdavimą į Extensible Markup Language (XML) kalbų pusę. Taip atsirado Chemical Markup Language (CML) [MR01]. Standartizavimas junginių lygyje buvo pradėtas dar anksčiau, sukūrus SMILES [Wei88] ir vėliau InChI [HMS+13] formatus.

Dabar dažnas šios srities mokslinis straipsnis turi papildomus duomenų CIF/CML dokumentus ir jau yra sprendimų [DDA+12], kurie „ropoja“ per viešai prieinamus straipsnius, renka normalizuotus dokumentus ir pateikia glaustą junginių informaciją pagal užklausą internetinės paieškos lauke.

Straipsniai, kurie neturi minėtų prisegtų dokumentų, yra indeksuojami kitais būdais. Straipsnių skaitmeninimas ir tolimesnis duomenų išgavimas yra sudėtingos užduotys, kurias skirtingose mokslo sferose sprendžia rankiniu, pusiau-rankiniu ([KKK+19]) arba automatiniu būdais. Kristalografiniai tyrimai ir jų rezultatai kol kas nėra išgaunami automatiškai, todėl šio proceso automatizavimas galėtų ženkliai padidinti mokslininkų darbo efektyvumą, su sąlygą, kad automatizuotas informacijos išgavimas turės pakankamai gerą tikslumą.

1. Darbo tikslas

Darbo tikslas yra pasiūlyti ir realizuoti naują sprendimą, gebantį išgauti kristalografinę junginių informaciją iš rišlių mokslinių tekstų. Šiam tikslui reikės surasti ir išanalizuoti kelis automatinio informacijos išgavimo būdus iš kristalografinių straipsnių, nepritaikytų kompiuteriniam skaitymui. Žmogui skirtą tekstą bus bandoma efektyviai apdoroti naudojantis įvairiais natūralios kalbos apdorojimo (angl. Natural Language Processing, NLP) metodais. NLP sritis yra aktyviai besivystanti ir kol kas negali pasiūlyti universalaus sprendimo įvairios formos ir tematikos tekstams, todėl skirtingų metodų efektyvumą reikės analizuoti kristalografinių tekstų kontekste.

Analizės pagrindu bus paruoštas programinis sprendimas, kuris turės numatytąjį (angl. default) darbo srautą (angl. workflow) bei galimybė modifikuoti teksto apdorojimo žingsnius pagal savo poreikį. Programinio sprendimo lankstumas tiesiogiai priklausys nuo sprendimo granuliacijos lygio, NLP susiformavusių standartų laikymosi ir teisingai suformuoto bendro įvesties/išvesties protokolo.

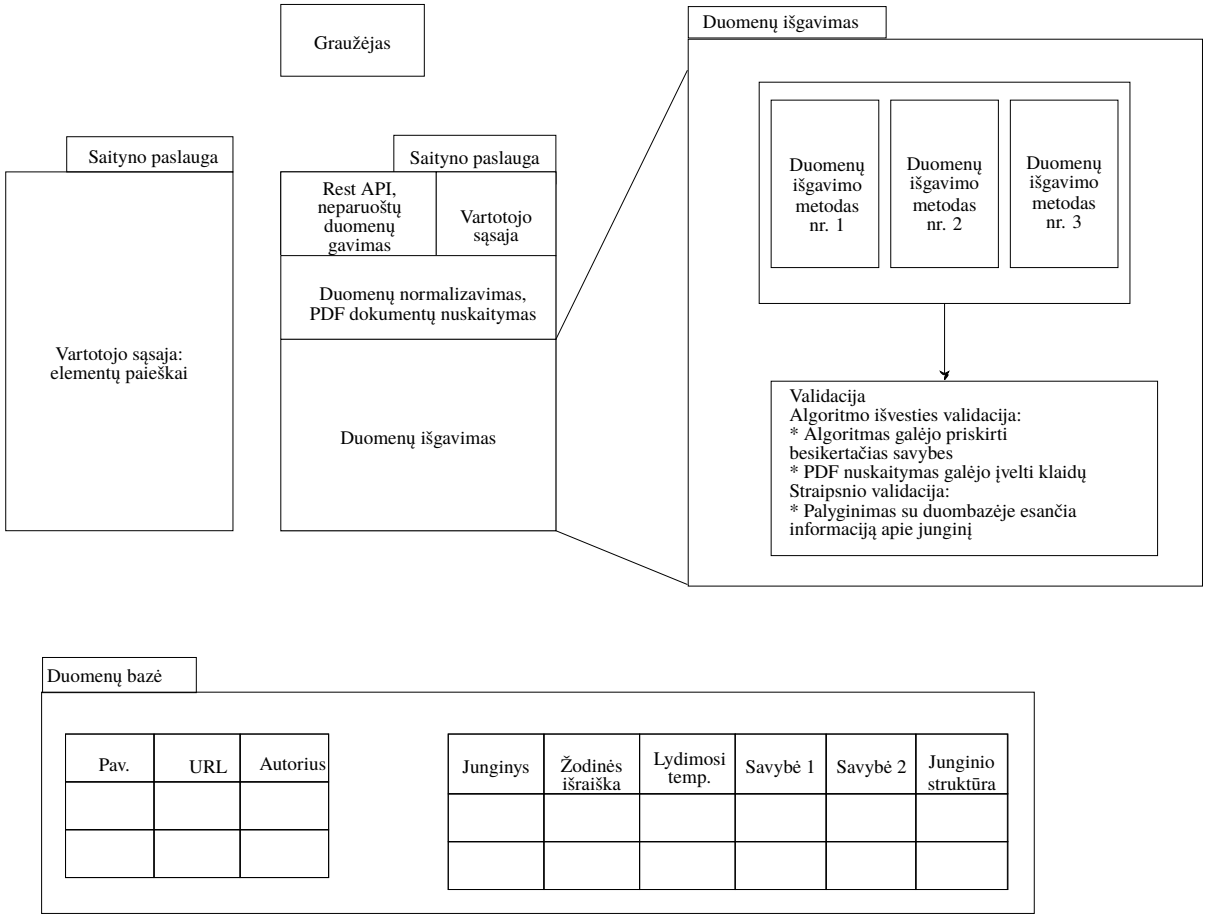
Įvairių netekstinių dokumentų nuskaitymas galėtų tapti atskiro tyrimo sritimi, todėl šio darbo ribose bus atliekamas tik teksto nuskaitymo žingsnis. Darbas fokusuosis į paruošto teksto apdorojimą. Svarbu pažymėti, kad informacijos išgavimas bus vykdomas tik iš tekstinio turinio. Lentelės ir grafikai taip pat nėra standartizuoti informacijos šaltiniai, todėl jie reikalauja atskiro dėmesio. Teksto analizės metodai šiems šaltiniams netiks, o tinkamo sprendimo paieška galėtų tapti atskiro tyrimo tikslu. Vienas iš galimų sunkumų yra aprašytų tekste junginių identifikavimas — tekste junginys gali būti paminėtas InChI/SMILES formatu, o lentelėje gali būti pateikta skaitinė nuoroda. Dažniausiai naudojamų formatų ir kristalografinio tekstų struktūros aiškinimasis taip pat tampa papildomu darbo tikslu.

2. Uždaviniai

Aukščiau suformuoti tikslai kelia moksliniam darbui šias užduotis:

- Taikytinų NLP metodų ir esamų sprendimų apžvalga.
  - Surasti ir išnagrinėti esamas metodikas cheminių tekstų apdorojimo sferoje. Rezultatu turėtų tapti esamų metodikų aprašymas, jų ypatumų apibendrinimas ir pakartotinio panaudojimo galimybių aprašymas.
  - Atskirų sprendimų analizė, sprendimų pranašumų trūkumų pasvėrimas padės suformuoti galutinio sprendimo architektūrą.
- Naujo sprendimo siūlymas ir įgyvendinimas.
  - Suformuoti sprendimo dizainą (struktūrą).
  - Sprendimo realizavimas.
- Gauto sprendimo vertinimas.
  - Suformuoti masyvą testavimui skirtų straipsnių, kurie struktūriškai būtų panašūs į tuos, iš kurių bus išgaunama informacija.
  - Rankiniu arba automatiniu būdu išgauti iš testuojamų straipsnių reikiamą informaciją. Ji taps pagrindu tikslumo ir atkūrimo analizei.
  - Ištirti kelis siūlomus informacijos išgavimo konvejerius su skirtingais parametrais, žodynais ir skirtingomis nuskaitymo taisyklėmis.
  - Įvertinti sprendimų pajėgumą tikslumo/atkūrimo (angl. precision/recall) metodu lyginant su rankiniu būdu išgauta informacija.

3. Sprendimo apžvalga



1 pav. Neformali sistemos schema

Darbo metu bus sukurtas sprendimas, kurio įvestis bus tekstiniai dokumentai, o išvestis — identifikuotų junginių lentelė su nurodytomis kristalografinėmis savybėmis. Šis sprendimas bus paruoštas tapti bendro informacijos išgavimo konvejerio dalimi. Potenciali vieta šiam duomenų išgavimo sprendimui yra viena iš pažymėtų ”Duomenų išgavimo metodo” blokų. Blokas nebus naudojamas su vaizdine informacija — tai bus duomenų normalizavimo, PDF nuskaitymo bloko dalimi.

Duomenų išgavimo blokas turės galimybę vykdyti informacijos išgavimo operacijas skirtingais metodais, pavyzdžiui, tekstiniai segmentai galės būti nuskaitomi skirtingais informacijos išgavimo srautais. Vaizdinės informacija bus nukreipiama į tam skirtus duomenų išgavimo metodus.

Labai svarbu yra užtikrinti tinkamą informacijos tikslumą. Klaidos gali pasitaikyti visuose teksto nagrinėjimo lygiuose — net PDF nuskaitymas gali palikti simbolių nuskaitymo klaidų.

Netikslumus turės identifikuoti validacijos blokas. Prieš informacijos suliejimą iš skirtingų srautų, validacijos blokas ieškos duomenų neatitikimų. Neatitikimo atveju bus spausdinama neatitinkanti informacija, kurią galės įvertinti naudotojas.

Validacija taip pat galės būti įvykdyta lyginant su patikrinta informacija. Blokas priims parametrus arba pavyzdžiui CIF dokumentą, kuriame bus aprašyta teisinga informacija. Taip pat,

blokas galės turėti ryšį su Crystallography Open Database ([GDM+23]) internetiniu resursu, kuris irgi gali būti laikomas patikrintos informacijos šaltiniu.

Galimybė įvesti teisingą informaciją padės įvertinti metodo/ų efektyvumą skaitmenine išraiška. Vertinant efektyvumą, informacijos tikslumas turės didesnį svorį už informacijos atkūrimą (angl. recall).

Sprendimas bus kuriamas atsižvelgiant į granuliarumo principus. Kiekvienas metodo žingsnis bus nepriklausomu įrankiu. Komunikacija tarp įrankių bus vykdoma tekstiniu formatu, kas leis nesunkiai patikrinti kiekvieno įrankio veikimą juodosios dėžės principu.

Bendrą sprendimo architektūrą nemažai įtakos esamų sprendimų analizė, bet jau dabar galima spėti, kad NLP veiksmai bus vykdomi Python aplinkoje. Duomenų rinkinių ruošimas ir apdorojimas galės būti vykdomas NodeJS aplinkoje. Technologinių apribojimų nenumatoma, išskyrus būtiną suderinamumą su Unix operacinėmis sistemomis.

4. Literatūros apžvalga

4.1. Problematika

Didžiausias darbo iššūkis yra teisingai interpretuoti žmogui skirtą tekstą skaitmeninėmis priemonėmis. Kadangi informacijos pateikimo būdų tekste yra begalinis skaičius, darbo ribose svarbiausia būtų sukurti teisingą informacijos nuskaitymą bent dažniausiai sutinkamose teksto struktūrose. Tokių struktūrų sąrašas įtraukia, bet nėra apribotas sekančių pateikimo variantų:

Reikiama informacija pateikta kintamojo priskyrimo principu

The structure of $Bi_2Al_4O_9$ has been refined anisotropically in the orthorhombic space group Pbam, with a = 7.7134(1), b = 8.1139(2), c = 5.6914(1)Å, and Z = 2.

2 pav. Gardelių parametrų pateikimas kintamųjų pagalba. Šaltinis: [CMG05]

Dažnai pasitaikantis gardelių parametrų pateikimo formatas. Tokiu būdų žymimų kintamųjų yra baigtinis skaičius, todėl paiešką bus galima atlikti reguliariųjų išraiškų pagalba. Žemiau yra pažymėtos tokiu būdų pateikiamų parametrų informacija.

Parametras	Reikšmė	Galimas žymėjimas	Matavimo vienetai
Gardelės a parametras	Realusis skaičius	a	Å - angstremai, nm
Gardelės b parametras	Realusis skaičius	b	Å - angstremai, nm
Gardelės c parametras	Realusis skaičius	c	Å - angstremai, nm
Gardelės alpha kampas	Realusis skaičius	α , a	laipsniai
Gardelės beta kampas	Realusis skaičius	β , b	laipsniai
Gardelės gamma kampas	Realusis skaičius	γ	laipsniai
Molekulių skaičius elementariajame narvelyje	Sveikas skaičius	Z	-

1 lentelė. Gardelių parametrai išreikšti kintamaisiais

Reikiama informacija pateikta laisva forma

$Sm_2(Fe,V)_{17+\delta}$ has a monoclinic structure with the space group P21/c, the same as $Nd_3(Fe,Ti)_{29}$, and with a periodicity based on 5 folds of the $CaCu_5$-type unit cell.

3 pav. Gardelių informacija pateikta laisva forma. Šaltinis: [KHO98]

Sakinyje (3) yra nurodyta $Sm_2(Fe,V)_{17+\delta}$ junginio struktūra ir erdvinė grupė (angl. space group). Sakinys taip pat yra įdomus tuo, kad yra minimas ir kitas junginys, kuris gali sukelti papildomų problemų informacijos nuskaitymo užduotyje ($Nd_3(Fe,Ti)_{29}$).

Faktinė informacija žymima keliuose sakiniuose, arba junginys sakinyje minimas netiesiogiai (pvz. The compound, former/latter compound, compound(1))

Sakiniuose su keliais minimais junginiais yra kritiškai svarbu teisingai nuskaityti semantinį ryšį. Užduotis pasunkėja, kai yra nurodoma į vieną iš junginių netiesiogiai (”pastarasis”, ”antras”,

”kitas” ir t.t.). Taip pat teiginys gali būti nepilnas vieno sakinio ribose, todėl paieška turės remtis informacija, surinkta iš prieš tai esančių sakinių.

Ne visi išvardinti informacijos pateikimo variantai gali būti paprastai nuskaityti vien tik reguliariųjų išraiškų pagalba. Laisva forma pateiktos informaciją nuskaitymas yra sudėtingas ir dažnai netikslus procesas. Deramai informacijos išgavimo kokybei pasiekti būtina suformuoti ieškomos informacijos ryšius su objektu.

Teksto apdorojimą gali sudaryti:

- teksto skaidymas leksemomis
- esybių paieška
- sakinio dalių žymėjimas
- sakinio struktūros nuskaitymas
- ieškomo simbolių diapazono žymėjimas tekste

Teksto apdorojimo rezultatas yra reikalingų atributų pažymėjimas, kurie bus naudojami tolimesniame faktų išgavime. Kituose skyriuose bus apžvelgti esami faktinės informacijos išgavimo sprendimai, žingsnių sekos ir jų teoriniai pagrindimai.

4.2. Informacijos išgavimas iš teksto

Teksto analitika (angl. text mining) yra plati mokslinė tema, kuri apima tikslingą informacijos išgavimą iš įvairaus tipo nestruktūruotų žmogaus parašytų tekstų [AZ12]. Bendras informacijos ir jos šaltinių kiekio augimas skatina mokslininkus ieškoti būdų, kaip apdoroti didelius tekstų kiekius ir taip išgauti vertingą informaciją. Teksto analitikos rezultatai priklauso nuo pirminių tikslų, bet bendros technikos dažnai būna panašios:

- teksto santrauka. Ilgas tekstas yra trumpinamas išlaikant bendrą teksto idėją ir svarbiausią informaciją.
- teksto dimensijų mažinimas.
- teksto klasifikavimas.
- esybių išgavimas.
- informacijos arba faktų išgavimas.

Greitas ir tikslus faktinės informacijos išgavimas iš nestruktūruotų tekstų tapo svarbia užduotimi eilėje sferų, kur tekstinių duomenų srautai turi dideles apimtis. Socialiniai tinklai, naujienų portalai, įvairūs verslai, mokslininkų bendruomenė kasdien generuoja labai didelius kiekius duomenų, kuriuos apdoroti rankiniu būdu tampa neįmanoma. Vieni pirmųjų bandymų kolektyviai spręsti šią problemą vyko varžybinėse mokslinėse konferencijose „Message Understanding Conference“ ([Sun91]). Tada daugelis dalyvavusių faktų išgavimo sistemų buvo paremtos apibrėžtomis taisyklėmis žodžiams ir žodžių grupėms (angl. concept nodes) [LCF+91]. Tokio tipo sistemos reikalavo didelių pastangų pildant sistemos žodyną ir identifikuojant visas galimas žodžių grupes. Taip pat taisyklės nebuvo universalios, todėl kiekvienam paieškos objektui turėjo būti koreguojamos.

```
(D-WORD DEAD
    :SYNTACTIC-TYPE SPECIAL-ADJECTIVE
    :SYNTACTIC-EXPECTATIONS
        (((assign *np-flag* t
                  *predicates* (append *predicates* (list *word*))
                  *part-of-speech* 'adjective
                  *cd-form* (make-special *word*)
                  *global-cn* *cd-form*)
          (next-packet
            ((test (eq *part-of-speech* 'noun)))
            ((test (eq *part-of-speech* 'adjective))))))
    :WORD-SENSES (dead1)
    :CN-DEFS ($LEFT-DEAD$ $FOUND-DEAD$ $FOUND-DEAD-PASS$))
```

4 pav. Žodžio „death“ apibrėžimas CIRCUS faktų išgavimo sistemoje [LCF+91]

```
(define-word $BOMBING-3$
    (CONCEPT-NODE
    ':NAME '$BOMBING-3$
    ':TIME-LIMIT 10
    ':SLOT-CONSTRAINTS '(((class organization *S*)(class terrorist *S*)
                          (class proper-name *S*)(class human *S*))
                         ((class phys-target *DO*) (class phys-target *PP*)))
    ':VARIABLE-SLOTS '(
                       ACTOR (*S* 1)
                       TARGET (*DO* 1 *PP* (is-prep? '(in))))
    ':CONSTANT-SLOTS '(
                       TYPE BOMBING))
    ':ENABLED-BY  '((active))))
```

5 pav. Žodžių grupės „bombing“ apibrėžimas CIRCUS faktų išgavimo sistemoje [LCF+91]

Vėliau taisyklių generavimą buvo bandoma automatizuoti, bet sistema reikalavo prižiūrinčio specialisto, kuris atmestų semantiškai nekorektiškas taisykles [LMS+93]. Sekantys žingsniai buvo orientuoti į informacijos nuskaitymo problemos skaldymą į kelias klasifikavimo problemas, kurias būtų galima spręsti žingsnis po žingsnio mašininio apmokymo metodais.

Svarbiausios tokio tipo užduotys faktų išgavimo procese yra esybių ir ryšio tarp sakinio žodžių nustatymas. Mažiau sudėtingas, bet taip pat labai svarbus yra teksto skaidymo leksemomis (angl. tokens) žingsnis.

4.2.1. Skaidymas

Teksto apdorojimas prasideda nuo tvarkingo teksto „skaldymo“ leksemomis (angl. tokenisation). Užduotis nėra paprasta net įprastiems bendro pobūdžio tekstams. Visu pirma tekstas dalinamas į sakinius, kur taškai ne visada reiškia sakinio pabaigą. Taškas gali reikšti sutrumpinimą, sveikosios dalies atskyrimą skaičiuje [RB21]. Cheminių tekstų kontekste taip pat dažnai sutinkami trupmeniniai skaičiai (pvz. β=95.776(8)), junginių žymėjimai (pvz. O–H·O).

Vėliau sakiniai skaidomi žodžiais. Tarpas šiam kontekste ne visada yra kokybiškas skirtukas, turint omeny, kad angliški tekstai kartais sujungia savarankiškus žodžius (pvz. white space/white-space), arba grupuoja žodžius brūkšneliais (pvz. the hold-him-back-and-drag-him-away maneuver). Į kelias šaknis ir morfologinę reikšmę turinčius priešdėlius ir priesagas gali būti atsižvelgta žodžių vektorizavimo žingsnyje (4.2.2). Po žodžių atskyrimo, jie dar gali būti normalizuoti. Po šio

žingsnio žodžiai gauna papildomą atributą, nurodantį bendrą žodžio formatą, pavyzdžiui, žodyninę formą arba žodžio šaknį. Žodyninės formos paieška gali būti sudėtinga dėl papildomos žodžių morfologinės analizės, todėl kartais yra naudojamas nesudėtingas šaknies atskyrimo algoritmas leksemoms išgauti [Por80].

Taip pat yra žodyno formavimo algoritmų, kurie neignoruoja jokių žodžių dalių. Šių algoritmų pagalba yra analizuojami tiksliniai tekstai ir jų pagrindu yra formuojami morfemų žodynai tolimesniam teksto apdorojimui. Pavyzdžiui, WordPiece ([WSC+16a]) algoritmas formuoja pirminį žodyną iš visų pateiktame tekste sutiktų simbolių. Vėliau algoritmas laipsniškai formuoja simbolių sekas, kurių sutikimo tikimybė yra aukštesnė, nei atskirų sekos komponentų. Tokiu būdu, tiksliniame tekste dažnai sutinkamos žodžių dalys (priešdėliai, šaknys ir kitos) galiausiai tampa žodyno dalimi.

4.2.2. Žodžių vektorizavimas

Kadangi mašininio apmokymo metodai operuoja su skaitinėmis išraiškomis, visi teksto žodžiai, arba morfemos, turi būti išreikšti atitinkamai. Žodžių išraiškai yra naudojami vektoriai. Vektorizavimo procesas siekia sukurti tokia vektorių sistemą, kurioje vektoriai atitinkantys žodžius atkartotų žodžių lingvistinį ryšį. Vektoriai apibūdinantys sinonimus turi būti greta, tuo tarpu antonimus reiškiantys vektoriai turi būti priešingi. Ryškiausiais tokių ryšių pavyzdys yra `vektorius(Vyras) + vektorius(Karalienė) = vektorius(Karalius) + vektorius(Moteris)`. Panašumas tarp vektorių yra skaičiuojamas skaliarine sandauga.

Paprasčiausi žodžių vektorizavimo būdai yra stàtiniai, arba išretinti (angl. sparse). Juose kiekviena vektoriaus komponentė atitinka žodį žodyne. Paprasčiausias pavyzdys — one-hot vector. Tokio vektoriaus reikšmės yra dvejetainės, kur vienetas nurodo konkretų žodį žodyne. Tankūs (angl. dense) vektorizavimo būdai yra kiek sudėtingesni. Tokio tipo vektoriai turi 50-1000 dimensijų, kurios neturi aiškios reikšmės. Tokio tipo vektorių erdvė gali būti sukurta naudojantis aibe skirtingų modelių. C&W, Continuous Bag of Words, Skip-gram, GloVe yra neuroninių kalbos modelių pavyzdžiai. Šie modeliai laikomi stàtiniais, nes tokie patys žodžiai sutiktì tekste gaus vieną ir tą patį vektorių. Priešingai negu stàtiniai, kontekstiniai vektorizavimo būdai generuoja vektorius atitinkančius žodžius priklausomai nuo žodžių eilės prieš ir po. Embeddings from Language Models (ELMo) yra vienas iš tokių modelių [PNI+18]. Jis išreiškia žodį pagal jį sudarančius simbolius konvoliucinio neuroninio tinklo (angl. Convolutional Neural Network) pagalba ir pagal žodžio kontekstą dvikrypčiu ilgos trumpalaikės atminties modeliu (angl. bidirectional Long Short Term Memory, BiLSTM).

Geriausius rezultatus daugelyje teksto apdorojimo užduočių siekia kontekstiniu Bidirectional Encoder Representations from Transformers (BERT) [DCL+18] vektorizavimu paremti sprendimai.

BERT modelių architektūra yra transformerių neuroninio tinklo (angl. Transformer Neural Network) tipo. TNN yra sudaryti iš koderio ir dekoderio. Abu yra sudaryti iš dėmesio mechanizmo (angl. self-attention) ir tiesioginio sklidimo sluoksnių. Kiekvieno sluoksnio lygyje, tarpinis žodžio-/morfemos vektorius tiesiogiai priklauso nuo dėmesio mechanizmo skirtų svorių kiekvienai įvesties

morfemai esančiai kontekste. Dėmesio mechanizmo tikslas – skaičiuojant kontekstinius morfemos vektorius skirti skirtingą svarbą viename kontekste esamoms morfemoms. Atskiros morfemos svarba nagrinėjamai morfemai išreiškiama svoriu. Svoris yra skaičiuojamas naudojant tris apmokytas svorių matricas W^Q, W^K, W^V. Vektorių sekai $(x_1,....x_n)$ matricų pagalba paskaičiuojami ”rakto”, ”užklausos” ir ”reikšmės” vektoriai:

$$q_i = W^Q x_i; k_i = W^K x_i; v_i = W^V x_i \text{ , kur } q \text{ - užklausos, } k \text{ - rakto, } v \text{ - reikšmės vektoriai} \quad (1)$$

Paverstas vektorius y_i yra visų pasvertų morfemų v_i vektorių suma:

$$y_i = a_{ij} v_j \quad (2)$$

α svoriai yra skaičiuojami taikant minkštojo maksimumo (angl. softmax) funkciją:

$$\alpha_{ij} = \frac{exp(\frac{q_i \cdot k_j}{\sqrt{d_k}})}{\Sigma_{k=1}^{n} exp(\frac{q_i \cdot k_z}{\sqrt{d_k}})} \quad (3)$$

Gauti svoriai yra pateikiami tiesioginio sklidimo sluoksnio įvesčiai. Kadangi šis vektorizavimo metodas nenumato rekurentinių tinklų naudojimo, jis atveria lygiagrečiu skaičiavimų galimybes ir bendrą apmokymo proceso paspartinimą. Viena iš modelio realizacijų (BERTbase [DCL+18]) naudoja 12 dėmesio mechanizmo ir tiesioginio sklidimo sluoksnių porų morfemų vektorizavimui. BERTbase į kontekstinius morfemų vektorius taip pat įtraukia ir pozicijos žymę, leidžiančią fiksuoti morfemų eilę.

BERTbase pirminiame modelyje (angl. pre-trained) morfemų vektoriai yra formuojami ”be mokytojo” apmokant modelį atlikti dvi užduotis:

- bandoma atspėti paslėptą morfemą pagal jos kontekstą taikant užmaskuotos kalbos modelį (angl. Masked Language Model, MLM).
- kito sakinio nuspėjimas (angl. Next Sentence Prediction, NSP). Modelis nuspėja tikimybę, kad du pateikti sakiniai prasmingai seką vienas paskui kitą.

BERTbase pirminis kalbos modelis yra apmokytas dideliu kiekiu literatūros anglų kalba (Book Corpus) ir Wikipedia straipsniais. Toks pirminis modelis sprendžia mokymosi perdavimo (angl. transfer learning) uždavinį. Pirminis TNN modelis gali būti papildomai mokomas konkretesnei teksto sferai arba pritaikytas visiškai naujai NLP užduočiai spręsti. BERTbase modelio įvestis – 512 skaitinių indeksų nurodančių į naudojamo žodyno įrašus. Didesnius morfemų kiekius modeliui galima pateiki pažingsniui. Morfemų žodynas BERTbase atveju buvo suformuotas Wordpiece [WSC+16b] algoritmo pagalba.

Pirminio, bendrai tekstų sričiai skirto, modelio vektoriai gali parodyti nepatenkinamus rezultatus taikant modelį specifinei tekstų sferai. Kalbos modelio neatitikimą tiksliniam tekstui sprendžia pritaikymo veiksmai (angl. fine-tune). Yra nemažai prieinamų modelių pritaikytų skirtingoms tekstų sferoms, pavyzdžiui, LegalBERT, TwHIN-BERT. Darbo kontekste galėtų būti pritai-

kytas BioBERT kalbos modelis, kuris yra papildomai apmokytas biomedicininiais tekstais: BioBERT ([LYK+19]) yra BERTbase kalbos modelis apmokytas PubMed santraukomis ir PMC straipsniais. Modelis gerina dalies cheminės tematikos NLP užduočių rezultatus, pavyzdžiui, BC5CDR ([LSJ+16]) cheminių junginių paieškos užduotį esybių žymėjimo sprendimas paremtas BioBERT kalbos modeliu rodo 93,68 tikslumo rodiklį, kuris nežymiai lenkia BERTbase rezultatą (P=90,94).

Taip pat yra modelių, kurie nesiremia mokymosi perdavimu — tokie modeliai yra pritaikomi konkrečiai teksto tematikai nuo pat žodyno formavimo žingsnio. SciBERT ([BLC19]) yra tokio modelio pavyzdys. Kitaip nei BioBERT, šis kalbos modelis buvo apmokomas tik moksliniais tekstais (82% biomedicininės tematikos tekstai ir 18% su informatiką susijusi literatūra). SciBERT kalbos modelis pritaikytas esybių žymėjimui, BC5CDR užduotį (simptomų ir cheminių junginių žymėjimas) atliko siekdamas 90,01 F1 statistikos rodiklį.

4.2.3. Esybių žymėjimas

Vienas iš svarbių žingsnių teksto apdorojimo užduotyje yra esybių žymėjimas (angl. named entity recognition). Žymės klasifikuoja žodį arba kelis žodžius kaip tam tikrą esybę, pavyzdžiui, žmogaus vardas, organizacijos pavadinimas, geografinė padėtis, telefono numeris. Šio darbo specifika verčia tekste fiksuoti vietas, kur yra minimi cheminiai junginiai. Tolimesniuose žingsniuose šios žymės leis susieti junginius su konkrečiomis skaitinėmis ar žodinėmis reikšmėmis.

Esybių žymėjimui yra naudojami 3 būdai: žymėjimas žodyno pagalba, žymėjimas pagal taisykles ir mašininiu mokymu pagrįsti sprendimai ([SJP+18]). Žodyninis esybių žymėjimo metodas yra apribotas savo žodyno turinio. Net mažiausi nesutapimai su žodyne esančiomis esybėmis neleis jų tinkamai pažymėti. Žymėjimo pagal taisykles metodas yra lankstesnis ir esybių klasifikacijai gali naudoti kontekstinę informaciją. Visgi šis metodas yra apribotas sukurtų taisyklių. Iš anksto nenumatyti atvejai nebus teisingai žymimi. Trečiasis mašininio mokymosi metodas yra lanksčiausias iš visų trijų, bet reikalauja didelio kiekio teisingai sužymėtų reikiamų esybių. Šiuo metu plačiai naudojamas metodas yra panašus į anksčiau pristatytą ELMo principą [MH16]. Pradžioje žodis simbolių lygyje yra vektorizuojamas konvoliucinių neuroninių tinklų pagalba. Vėliau vektorius yra papildomas kontekstine informacija aplink žodį išgauta dvipusiu ilgu laikinosios atminties tinklu. Pabaigoje gauti vektoriai papildomai yra perduodami į sąlyginiais atsitiktiniais laukais (angl. Conditional Random Fields, CRFs) paremtą tinklą, kuris leidžia fiksuoti esybės žodžius ne kaip nepriklausomus vienetus, o kaip esybę sudarančią žodžių seka. Biologiniuose tekstuose vieni geriausių rezultatų yra gaunami BERT vektorizavimo metodais, papildomai ištreniruotais PubMed duomenų bazės straipsniais. Biologiniams tekstams pritaikytas kalbos modelis, pavyzdžiui BioBERT, yra papildomai pritaikomas spręsti klasifikavimo problemą. Tokiam tikslui BERT architektūros modeliai yra pritaikomi pridedant papildomą pilnai prijungta paslėptą sluoksnį su dvejomis išvestimis: ieškomo diapazono pradžios ir pabaigos vektoriai [SS19].

Skirtingų sprendimų tikslumas yra vertinamas esybių nuskaitymo užduočių ribose. Viena tokių yra BioCreative V Chemical Disease Relation (BC5CDR) [LSJ+16]. Užduotis pateikia 1500 anotuotų PubMed duomenų bazės straipsnių, kuriuose pažymėti ryšiai tarp cheminių junginių ir jų

sukeliamų simptomų. Anotuotų cheminių junginių skaičius yra lygus 15411. 500 anotuotų straipsnių modelio treniravimui, 500 tarpiniam modelio tikrinimui (angl. development) ir likę 500 modelio tikslumo nustatymui. 2019 metais BC5CDR užduoties ribose geriausiai cheminius junginius žymėjo BERT principu veikiantis modelis, kuris pasiekė 93,47 F1 statistikos rodiklį [LYK+19]. F1 statistika arba F_1 yra universalus testavimo įvertis, kuris pateikia sprendimo tikslumą ir atkūrimo lygį viena skaitine išraiška (6).

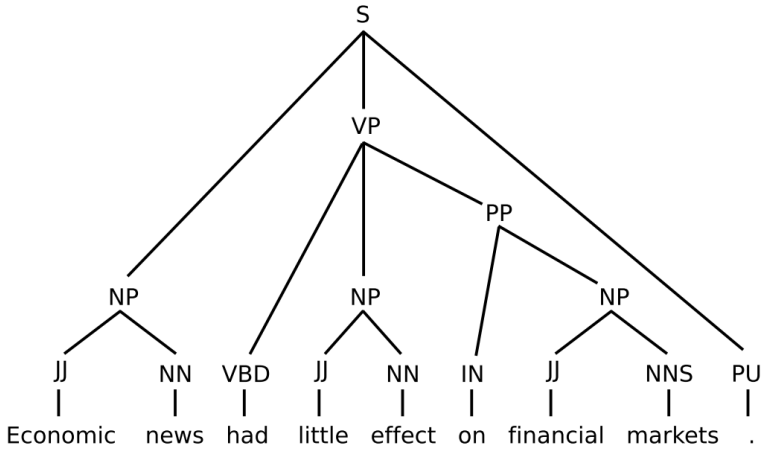
$$P = \frac{TP}{TP + FP} \text{ , kur P - tikslumas, TP - teisingai teigiami, FP - klaidingai teigiami} \tag{4}$$

$$R = \frac{TP}{TP + FN} \text{ , kur R - atkūrimas, TP - teisingai teigiami, FN - klaidingai neigiami} \tag{5}$$

$$F_1 = \frac{2 * P * R}{P + R} \tag{6}$$

4.2.4. Žodžių ryšių nuskaitymas

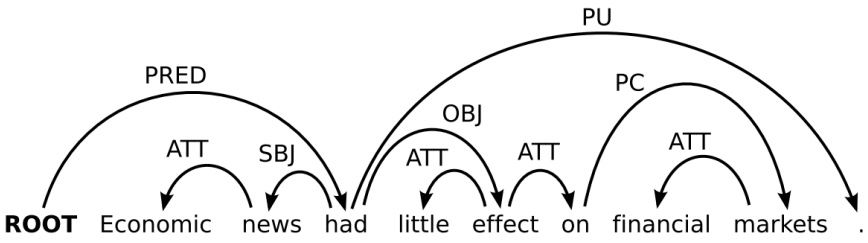
Žodžių vektorizavimas nurodo mašinai žodžių rinkinį, iš kurių yra sudarytas sakinys, bet ryšys tarp jų turi būti fiksuojamas atskirame žingsnyje. Sakinio žodžių ryšys gali būti fiksuojamas iš dviejų skirtingų perspektyvų: frazinių struktūrų ir žodžių priklausomybių. Frazių struktūrų gramatikos (angl. phrase structure grammar) žymėjimas gali pasirodyti natūralesniu iš lingvistinės pusės. Žodžiai yra grupuojami į frazes, kurios įgauna struktūrinį identifikatorių, pavyzdžiui veiksmažodžio frazė (angl. verb phrase, VP), daiktavardžio frazė (angl. noun phrase, NP).



6 pav. Sakinio *Economic news had little effect on financial markets.* frazinių struktūrų vizualizacija [MN11]

Antrasis, priklausomybių nuskaitymo (angl. dependency parsing) būdas yra kiek kompaktiškesnis. Jis leidžia nustatyti sakinio šaknį ir priklausomus nuo jo žodžius nepriklausomai nuo

žodžių eilės tvarkos.



7 pav. Sakinio *Economic news had little effect on financial markets.* žodžių priklausomybių vizualizacija [MN11]

Priklausomybių nuskaitymas yra atliekamas priklausomybių gramatikos pagrindu. Tokio tipo gramatikų buvo sukurta nemažai (Functional Generative Description, Meaning-Text Theory, and Word Grammar). Visos gramatikos turi bendrą tikslą — pasiūlyti sintaksinę struktūrą, kurioje žodžiai yra susieti vienapusiu priklausomybės santykiu. Visos minėtos gramatikos identifikuoja tradicines gramatines struktūras (veiksnius, tarinius, papildinius (angl. subject, predicate, object)), bet kiti mažiau svarbūs ryšiai yra interpretuojami skirtingai. Nors ir abu sintaksinio nuskaitymo principai laikomi tarpusavyje nesuderinamais ir nekonvertuojamais iš vieno į kitą, yra sintaksinio nuskaitymo teorijų, kurios vienu metu naudoja abu sakinio nuskaitymo principus.

Sakinio struktūros žymėjimo algoritmai yra skirstomi į duomenimis (angl. data-driven) ir gramatika pagrįstus (angl. grammar-based). Duomenų paradigma veikiantys sprendimai nustato žodžių priklausomybes naudodami prieš tai apmokytus modelius. Tokie sprendimai šiuo metu yra populiariausi dėl nesudėtingo pritaikymo bet kuriai kalbai, kuri turi pavyzdines priklausomybių struktūras modelio treniravimui. Taip pat, populiarumą paskatino jau sužymėtų struktūrų skirtingomis kalbomis publikavimas CoNLL užduotyje [MN11]. Duomenimis pagrįsti metodai skirstomi į dvi grupes: paremti perėjimais (angl. transition-based) ir paremti grafais (angl. graph-based). Pereinamuoju principu veikiantys sprendimai priklausomybes nustato po vieną. Priklausomybės klasifikavimo modeliai palaipsniui suformuoja sakinių priklausomybes vaizduojančius medžius. Modelio įvestį gali sudaryti pats žodis (vektoriaus išraiška), jo standartinė išraiška, kalbos dalies identifikatorius arba kitos lingvistinės žodžio savybės, išgautos prieš nustatant žodžių priklausomybes. Modeliai yra treniruojami rankomis sužymėtų medžių struktūrų (angl. treebank) pagalba. Tuo tarpu grafais paremtuose sprendimuose tinkamiausia sakinio priklausomybių struktūra surandama visų galimų variantų aibėje. Optimalus variantas yra nustatomas pagal visų grafo viršūnių įvertinimų sumą. Šis metodas yra tikslesnis ilguose sakiniuose, kur tarpai tarp priklausomų žodžių yra vidutiniškai ilgesni.

Universal Dependencies ([NDG+16]) yra vienas iš projektų, kurie siūlo savo žodžių priklausomybių modelį, tinkantį daugeliui pasaulio kalbų. Universal Dependencies bendruomenės projekto tikslas žymėti įvairiomis kalbomis parašytų tekstų struktūras ir kitus sakinio/žodžio lygio atributus ir pateikti juos kompiuteriu skaitomu formatu (CoNLL-U). Universal Dependencies išskiria priklausomybes surištas su tariniu (angl. Clausal Argument Relations) ir tikslinančias priklausomybes (angl. Nominal Modifier Relations).

Dauguma sakinio struktūrų nuskaitymo sprendimų taip pat pažymi sakinio dalis ([Str18], [ABB+19]), bet yra ir sprendimų, priklausančių nuo kitų žodžių atvaizdavimo būdų (pavyzdžiui, skiemenų vektorių [JCP+]).

Sakinio struktūros nuskaitymo tikslumą parodo Labeled Attachment Score (LAS) ir Unlabeled Attachment Score (UAS) koeficientai. UAS nurodo žodžių proporciją, kuriems buvo nustatytas teisingas valdantis žodis. Tuo tarpu LAS nurodo žodžių procentą, kuriems buvo teisingai priskirtas ir valdantis žodis, ir priklausomybės tipas.

Du iš apžvelgtų sprendimų geba žymėti sakinio struktūrą. UDPipe [Str18] CoNLL 2018 Universal Dependencies užduotį atliko UAS — 86,8, LAS — 83,15 tikslumu. BERT modelis pritaikytas atlikti klasifikacijos (UD žodžių žymėjimui) ir priklausomybių nuskaitymo užduotis parodė UAS=91,55, LAS=89,06 rezultatą. SciSpacy ([NKB+19]) `en_core_sci_scibert` modelis parodė UAS — 92,03 ir LAS 90,25 rezultatus mokslinių tekstų rinkinyje GENIA. [KOT+03]

Nuskaityti žodžių ryšiai galėtų būti vėliau ieškomi pagal iš anksto paruoštų taisyklių sąrašą. Tokio tipo taisykles galima apibrėžti Semgrex įrankiu [CCG+07]. Jis buvo sukurtas reikiamų struktūrų paieškai semantinių priklausomybių medžiuose. Paieška jo pagalba yra vykdoma pagal paieškos užklausas, kurių sintaksė yra panaši į kitus lingvistinių medžių struktūrų paieškos įrankius (tgrep/Tregex). Sintaksė leidžia apibrėžti ieškomų medžio viršūnių savybes, jas siejančias priklausomybes ir jų kryptį. Pateiktas faktinės informacijos išgavimo principas gali duoti reikiamus rezultatus tik su prielaida, kad prieš tai atlikti žingsniai buvo atlikti teisingai. Netikslumai teksto nuskaitymo, skaidymo žodžiais metu tiesiogiai įtakoja kalbos dalių, priklausomybių žymėjimą ir taip pat esybių paiešką. Kadangi išvardinti atributai tampa paieškos orientyrais, konvejerio tikslumo analizė taps kritiškai svarbia užduotimi darbo pabaigoje.

4.2.5. Klausimų-atsakymų užduotis

Kitas galimas būdas susieti reikšmes su dominančiu kontekstu yra klausimų-atsakymų modelių taikymas. Klausimų-atsakymų užduotis plačiąją prasme gali būti susieta su sudėtingomis informacijos paieškos (angl. information retrieval) sistemomis. Sistemos grąžinami dokumentai vėliau yra apdorojami galutinio atsakymo pateikimui. Kontekstinės informacijos apdorojimas pateikiant atsakymą yra skirstomas į du tipus: generuojamo atsakymo ir teksto išgavimo. Teksto išgavimas šiam darbui yra labiau aktualus, kadangi išgauta informacija nebus toliau interpretuojama. Teksto išgavimo užduotį NLP kontekste, panašiai kaip esybių paiešką, galima apibrėžti kaip teksto diapazono klasifikavimo (angl. span classification) užduotį. BERT architektūros modeliuose tokią problemą sprendžia tiesinis sluoksnis modelio gale. Pateikus modeliui klausimą ir kontekstą, atskirtus specialiu skirtuku, paskutinis linijinis sluoksnis yra apmokomas duomenų rinkiniu, kad galėtų nustatyti labiausiai tikėtinus vektorius ieškomo teksto diapazono pradžiai ir pabaigai.

Šiam tikslui dažniausiai yra naudojami bendros tematikos duomenų rinkiniai sukurti SQuAD (angl. Stanford Question Answering Dataset) NLP užduoties ribose. Prieinamos yra dvi SQuAD duomenų rinkinio versijos. Antroji buvo papildyta klausimais ir kontekstais, kuriuose atsakymas nefigūruoja. BERT architektūros modelis pritaikytas teksto diapazonų žymėjimui rodo 93,2 F1 sta-

tistikos rodiklį atliekant SQuAD1.1 užduotį ir 83,1 rodo SQuAD2.0 atveju [DCL+19]. Geri rezultatai atliekant teksto diapazono žymėjimo užduotį nurodo, kad teoriškai galima spręsti informacijos išgavimo problemą laipsniškai tikslinant klausimus. Tokiu būdu klausime galėtų būti fiksuojamos ir antrinės sąlygos, pavyzdžiui, temperatūra. Tokio modelio tikslumas tiesiogiai priklausytų nuo paruošto duomenų rinkinio kokybės.

5. Analitinė dalis

Šiame skyriuje bus apibrėžtos duomenų išgavimo sistemos gairės, jos įvestys ir išvestys, sistemos vertinimo kriterijai bei pavyzdinių duomenų šaltiniai. Bus paminėti galimi iššūkiai ruošiant sistemą.

5.1. Sistemos įvestys ir išvestys

Svarbu yra suskaidyti visą teksto apdorojimo seką į atskirus, nepriklausomai tikrinamus žingsnius. Pradžioje PDF formato dokumentai turėtų būti nuskaityti kompiuterinės regos (angl. computer vision) pagalba rastrinių dokumentų atveju arba sintaksiniais analizatoriais (angl. parser). PDF formato dokumentų nuskaitymo užduotis nėra triviali (ypač rastrinių dokumentų atveju), kadangi dokumentų formatavimas gali skirtis kiekviename moksliniame žurnale. Senesnių straipsnių skenuotos kopijos taip pat gali turėti nuskaitymą trikdančius artefaktus. Teksto sekos nuskaitymą apsunkina ir žodžių kėlimas į kitą eilutę. Atskiru nuskaitymo žingsniu taip pat galėtų tapti lentelių nuskaitymas – daugelyje dokumentų jos yra skirtingai formatuotos. Nors būtent lentelėse dažniausiai yra atvaizduojami kristalografiniai parametrai, visgi nemaža dalis tyrimais nustatytų parametrų lieka autoriaus parašytuose tekstuose. Apibendrinant – teksto ir lentelių nuskaitymas iš PDF formato dokumentų yra užduotis verta atskiro dėmesio. Šio darbo kontekste bus siekiama išgauti kristalografinius parametrus iš jau nuskaityto teksto. Šio darbo sėkmė pagrįstų vėlesnę PDF nuskaitymo sprendimo paiešką.

Aprašomos sistemos įvestimi bus UTF-8 teksto paketai. Išvestis – nuskaitytų duomenų sekos $\{junginio\_pavadinimas, parametro\_pavadinimas, reiksme\}$ paruoštos įterpimui į Crystallography Open Database. Atskyrus PDF dokumentų nuskaitymą nuo sistemos, pagrindiniu darbo iššūkiu tampa laisva forma parašytų tekstų apdorojimas, kuriuose informacija yra pateikiama daugeliu skirtingų formų. Junginiai, kaip ir kristalografinė informacija, gali būti išreikšti keliomis notacijomis. Sistema turi užtikrinti išgautos informacijos tikslumą, leidžianti rinkti informaciją pusiau automatiniu būdu su minimaliu žmogaus įsikišimu.

5.2. Tekstų šaltiniai

Kristalografinių tekstų bus ieškoma atvirai prieinamuose kristalografiniuose mokslo žurnaluose tiesiogiai (pvz. Acta Crystallographica, Journal of the American Chemical Society) ir mokslinių straipsnių talpyklose (pvz. arXiv.org). Prioritetas bus teikiamas straipsniams, tiriantiems mažų molekulių (iki 10 kDa) kristalografines savybes. Baltymų kristalografiniai aprašymai taip pat dalinai galėtų tikti modelių apmokymui (pvz. cheminių junginių paieškai), bet šių veiksmų bus imamasi, jeigu pritrūktų laisvai prieinamų straipsnių kristalografine tematika.

Kitas svarbus reikalavimas kristalografinių tekstų šaltiniui yra formatuotų dokumentų prieinamumas ir galimybė juos nuskaityti automatiniu būdu. Tinkami dokumentų atvaizdavimo variantai galėtų būti:

HTML formatas. Dažnai HTML (angl. HyperText Markup Language) dokumentai publikuojami viename šaltinyje išlaiko vienoda struktūrą (pvz. Acta Crystallographica E, talpinantis beveik 45 tūkstančius straipsnių). Tai leidžia naudoti vieną teksto išgavimo algoritmą visiems šaltinio straipsniams. Taikytis prie skirtingų šaltinių tektų mažiau, jeigu straipsniai būtų publikuojami griežtai aprašytų formatu (pvz. Scholarly HTML), bet tokio tipo formatai nėra plačiai naudojami.

XML dokumentai. Kadangi formatas yra skirtas tolimesniam jo pritaikymo sukonkretinimui (schemos arba Document Type Definition (DTD) pagalba), šaltiniai gali formatuoti dokumentus sau palankiu būdu. Šiam formatui užtektų vieno teksto išgavimo algoritmo visiems šaltinio dokumentams nuskaityti. XML formatavimą naudoja EuropePMC talpykla, turinti daugiau 40-ies milijonų dokumentų.

LaTeX dokumentai. Kartais moksliniai straipsniai yra publikuojami kartu su pirminiu LaTeX formato dokumentu (pvz. arXiv talpykla, turinti daugiau dviejų milijonų straipsnių). Tokio tipo dokumentai taip pat yra tinkami, kadangi tekstą iš jų galima išgauti LaTeX dokumentus į tekstą verčiančiais įrankiais.

5.3. Sistemos vertinimas

Informacijos išgavimo kokybę galima įvertinti paruošus tikslinį, panaudojimą atvaizduojantį, tikrinimo duomenų rinkinį (angl. validation dataset) sudarytą iš įvesčių (kristalografinių tekstų) ir išvesties sekų ($\{junginio\_pavadinimas, parametro\_pavadinimas, reiksme\}$), paruoštų rankiniu arba automatiniu būdu. Išgautos informacijos kokybės vertinimo kriterijai turi atitikti prioritetus, taikomus darbui. Kadangi surinkta informacija galiausiai bus įterpta į kristalografinę duomenų bazę, reikalavimai tikslumui yra svaresni, nei atkūrimui.

Taikomus išgautos informacijos vertinimo rodiklius galima skirstyti į du tipus:

Lingvistinius. Dažniausiai naudojami klausimų-atsakymų užduoties vertinimui. Daugeliu NLP taikymų atvejų išvestis yra pateikiama naudotojui tolimesnei interpretacijai, todėl vertinimo rodikliais yra bandoma užfiksuoti ne tik visiškai tikslius atsakymus, bet ir išvesties atsakymo sankirtą su tikruoju atsakymu. Toks metodas skaičiuoja tikslumą ir atkūrimą kiekvienam sistemos spėjimui, paversdamas spėjimus ir teisingus atsakymus morfemų masyvais ir ieškodamas tarp jų atitikimų. Tokiu būdu tikslumas, atkūrimas ir F1 statistika gali būti skaičiuojami kiekvienam spėjimui atskirai. Galutinis vertinimas yra visų gautų atskirų rodiklių vidurkis (angl. macro-averaged) [RZL+16].

Faktinius. Šis vertinimo tipas klausimų-atsakymų sistemų kokybės vertinimui yra naudojamas rečiau, kadangi jis griežtai tikrina atitikimus su tikrinimo duomenų rinkiniu. Šis metodas gali turėti švelninančias sąlygas, pavyzdžiui, artikeliai, raidžių registras arba tarpai gali būti ignoruojami galutiniame vertinime. Tokiu būdu, ilgesni spėjimai, kurie taip pat pamini ir teisingą atsakymą, nėra teigiamai fiksuojami galutiniame rodiklyje. Užduoties atlikimo tikslumas ir atkūrimas yra skaičiuojami bendrai visų atsakymu išvesčiai. Taip pat prie faktinių rodiklių galima priskirti ir tikslių atitikimų (angl. exact match) santykį.

Visos kristalografinės informacijos sistemos ir jos atskirų komponentų vertinimo metu buvo nuspręsta skirti daugiau dėmesio faktiniams vertinimo kriterijams, kadangi Crystallography Open Database duomenų bazės pildymas kelia aukštus reikalavimus duomenų tikslumui. Aukščiau aprašyti vertinimo kriterijai gali būti taikomi ne tik visos sekos vertinimui, bet ir atskirų informacijos išgavimo žingsnių tikrinimui. Kadangi kristalografinės informacijos išgavimas iš mokslinių straipsnių yra sudėtinga užduotis, reikalaujanti skaidymo į žingsnius, kiekvienas žingsnis bus kuriamas ir vertinamas atskirai.

6. Sistemos projektavimas

Šiame skyriuje pristatysiu komponentus, iš kurių bus sudaryta kristalografinės informacijos išgavimo sistema. Bus palygintos teksto apdorojimo sekų alternatyvos ir pateiktas tolimesnių veiksmų planas.

Modernūs NLP metodai numato bendresnių sistemų laipsnišką pritaikymą tikslinėms siauroms užduočių sferoms. Pirminis modelis (angl. pre-trained model) atlieka bazinės kalbos modelio vaidmenį. Literatūros apžvalgoje minėti pirminiai modeliai būna apmokyti didelių tekstų kiekiu ir šie kalbos modeliai gali būti apmokomi papildomai tam tikrai tekstų tematikai (angl. fine-tuning). BERT architektūros atveju, gauta kalbos modelio įvestis yra n morfemų identifikatorių, kurie nurodo į morfemas modelio žodyne, o išvestis – n vektoriai nusakantys (angl. embeddings) įvesties morfemas pateiktame kontekste (n - modelio palaikomas įvesties ilgis). Kontekstais toliau bus vadinami įvesties tekstai.

Vėliau kalbos modelių architektūrą galima papildyti linijiniu sluoksniu, kuris yra apmokomas atlikti klasifikacijos užduotį. Vieno žingsnio klasifikavimą galima būtų taikyti, jeigu pateiktuose tekstuose būtų minimas tik vienas cheminis junginys ir su juo siejami kristalografiniai parametrai. Tačiau tekstuose cheminių junginių kiekis nėra ribotas, todėl svarbu yra susieti cheminius junginius su būtent jiems priklausančiais kristalografiniais parametrais.

Taigi, problemą galimą išskaidyti į dvi dalis: cheminių junginių identifikavimą ir konkretaus cheminio junginio parametrų paiešką.

6.1. Cheminių junginių identifikavimas

Cheminių junginių identifikavimas yra viena iš esybių žymėjimo pritaikymo sferų. Tai yra klasifikavimo užduotis, kuri pažymi įvesties teksto morfemas žymėmis. Modeliai apmokyti literatūros apžvalgoje minėtu BC5CDR duomenų rinkiniu žymi dalį cheminių (molekulinių) formulių ir junginių pavadinimų, bet klysta žymėdami šias junginių atvaizdavimo formas:

- liekanas žyminti cheminė formulė (pvz. [Fe(C_5H_5)($C_30H_32O_7P$)][ŠLC09]).
- IUPAC (International Union of Pure and Applied Chemists) organinių junginių nomenklatūra (pvz. 1,2:5,6-Di-O-isopropylidene-α-D-3-glucofuranosyl (Rp)-2-(diphenylphosphino)ferrocene-1-carboxylate [ŠLC09]).
- dalinės (angl. moiety) formulės (pvz. $Na_2SeO_4 \cdot 1.5H_2O$ [For15]).

Buvo suplanuota šių atvaizdavimo formų duomenų rinkinį, kurio pagrindu bus papildomai apmokyti BC5CDR užduotį atliekantys modeliai.

6.2. Junginio parametrų išgavimas

Išgavusi junginių pavadinimus, sistema galėtų atlikti su junginiais susijusių parametrų paiešką. Junginio ryšį su parametru galima nustatyti priklausomybių nuskaitymu ir sakinių schemų atitikimų paieška arba klausimų-atsakymų modeliais paremtais sprendimais.

6.2.1. Priklausomybių nuskaitymas ir tolimesnė paieška

Toks būdas apima žodžių priklausomybių žymėjimą ir paieškos atlikimą aprašytą literatūros apžvalgoje. Šio sprendimo sėkmė tiesiogiai priklauso nuo žodžių priklausomybių žymėjimo kokybės. Kadangi kristalografiniai tekstai skiriasi nuo Universal Dependencies duomenų rinkinyje esamų tekstų stiliaus, priklausomybių žymėjimo modelius reikėtų papildomai pritaikyti. Priklausomybių modelio apmokymas ”be mokytojo” kol kas nėra kokybiškai įgyvendinamas – modeliui apmokyti yra būtina paruošti duomenų rinkinį su tinkamu priklausomybių žymėjimu. Universal Dependencies žymėjimo modelis turi sudėtingas, griežtai aprašytas priklausomybių medžio formavimo taisykles, todėl tokio duomenų rinkinio ruošimas pareikalautų per daug pastangų. Taip pat tektų skirti laiko priklausomybių paieškos optimizavimui, pavyzdžiui, optimaliai žymėti sakinius, kurie galėtų turėti informaciją susijusią su ieškomu parametru. Pirminė sakinių paieška galėtų būti įgyvendinta reguliariųjų išraiškų pagalba, klausimų-atsakymų arba teksto santraukos modeliais. Galima spėti, kad turint priklausomybių nuskaitymo sprendimą, kuris veiktų aukštu tikslumu, galima būtų pradėti ruošti parametrų paieškos taisykles. Kadangi sakinių struktūra moksliniuose straipsniuose nėra apibrėžta (bendras yra tik akademinis kalbos stilius), tokių parametrų išgavimo taisyklių gali būti neribotas kiekis. Taisyklių ruošimas pareikalautų per daug darbo, tačiau galutinio sprendimo rezultatai būtų gerai nuspėjami ir parametrų neišgavimo priežastys galėtų būti tiksliai fiksuojamos.

6.2.2. Parametrų išgavimas klasifikuojant teksto morfemas

Šis parametrų išgavimo būdas galėtų radikaliai supaprastinti visą junginių-parametrų porų išgavimo sprendimą iki vieno modelio, gebančio klasifikuoti junginį kartu su visais dominančius junginio parametrais. Tokio sprendimo vykdymo laikas ženkliai lenktų alternatyvas dėl savo paprastumo — kitaip nei klausimų-atsakymų modelio atveju, klasifikavimo procesą užtektų atlikti viena kartą šimtams morfemų ir tokiu būdu junginiai ir jų parametrai įgytų diapazonų žymes.

Tačiau toks sprendimas reikalautų papildomos euristikos atvejams, kada kontekste yra aiškinami keli junginiai. Yra darbų ([GCT+23]), kurie šiai problemai spręsti pritaikė BERT pagrindu veikiantį teksto klasifikavimo modelį. Klasifikavimas vyko pateikiant modeliui po vieną sakinį, kuriame vieno junginio ir vieno parametro (Curie temperatūros) diapazonai buvo pažymėti specialiomis žymėmis. Taigi, modelis buvo vykdomas kiekvienai įmanomai junginio-temperatūros porai sakinyje, o dvejetainė modelio išvestis nurodė pateiktos junginio-temperatūros poros ryšį arba jo nebuvimą. Straipsnyje aprašytas ryšių identifikavimo klasifikatorius sakinio lygyje parodė 0,72 tikslumo ir 0,64 atkūrimo rodiklius (0,68 F1 statistika). Nors šis sprendimas apibrėžia ir izoliuoja ryšių išgavimo problemą, konvejerio unikalumas kelia aukštus reikalavimus apmokymo duomenų rinkiniams. Pirminis junginių ir parametrų klasifikavimas reikalauja kruopščiai anotuoto duomenų rinkinio – klasifikavimo modelis žymi visas įvesties morfemas, todėl anotacijos turi būti tikslios ir išsamios. Klausimų-atsakymų modelio apmokymas yra paprastesnis dėl galimybės praleisti dalį kontekste esančių klausimų-atsakymų porų. Šiuo metu yra viešai prieinamų kokybiškų, rankomis paruoštų duomenų rinkinių, žyminčių cheminius junginius, tačiau kol kas nėra viešų kristalografi-

nių parametrus klasifikuojančių duomenų rinkinių. Ryšių klasifikavimas taip pat reikalauja unikalaus duomenų rinkinio.

6.2.3. Klausimų-atsakymų modelis

Kitas būdas išgauti su cheminiais junginiais susijusius parametrus gali būti paremtas klausimų-atsakymų modeliu. Atributų reikšmių (angl. attribute-value) išgavimo užduotis jau buvo sprendžiama klausimų-atsakymų modeliais produktų aprašymams apdoroti ([WYK+20]). Toks sprendimas demonstruoja lankstumą ir pritaikymo srities plėtimo galimybes: vienas modelis geba atlikti kaip modeliui matytų, taip ir nematytų atributų ir jų reikšmių paiešką. Modelį apmokius pakankamu kiekiu klausimų apie atributų reikšmes, modelis suformuoja abstrakčius atributo ir jo reikšmės ryšius pateiktuose kontekstuose, ir todėl, remiantis šiais ryšiais, atsiveria galimybės klausti modelio unikalių atributų reikšmių.

Klausimų-atsakymų modelį galima apmokyti paruoštu parametrų išgavimo duomenų rinkiniu. Duomenų rinkinio ruošimas yra kruopštus ir daug resursų reikalaujantis darbas, todėl sunku tikėtis, kad pavyks suformuoti daugiau nei 10000 teksto pastraipų turintį duomenų rinkinį. Pradėti galima nuo paskutinio sluoksnio apmokymo SQuAD2.0[RJL18] duomenų rinkiniu, kuris talpina daugiau 100000 klausimų-atsakymų įrašų. Apmokius modelį vien šiuo duomenų rinkiniu, modelis jau kartkartėmis galės nuspėti parametro reikšmės vietą. Šio darbo kontekste svarbu atkreipti dėmesį būtent į 2.0 duomenų rinkinio versiją, kadangi jis buvo papildytas klausimais, kurių kontekstai neturi atsakymo.

6.3. Planas

- Parametrų išgavimo komponento realizacija.
  - Suformuoti apmokymo ir tikrinimo duomenų rinkinius, ”a”, ”b”, ”c” gardelės parametrams ir simetrijos grupių išgavimui.
  - Įvertinti BERT architektūros modelius apmokytus SQuAD2.0 duomenų rinkiniu.
  - Paruošto duomenų rinkinio pagalba apmokyti BERTbase, BioBERT, SciBERT kalbos modelius, ištirti modelius apmokytus skirtingas duomenų rinkinių variacijomis.
  - Papildyti duomenų rinkinį duomenimis iš daugiau šaltinių ir pridėti kitų gardelės parametrų pavyzdžių.
- Cheminių junginių išgavimo komponento realizacija.
  - Suformuoti duomenų rinkinį cheminių formulių ir junginių pavadinimų identifikavimui.
  - Apmokyti modelį esybių žymėjimo modelį BioBERT ir SciBERT pagrindu.
- Komponentų sujungimas ir visos sekos tikslumo tikrinimas. Baigiamasi apdorojimas (angl. post-processing).

7. Sistemos realizacija

Šiame skyriuje bus aprašyta sistemos komponentų realizacija ir tolimesnis atskirų komponentų sujungimas. Pradžioje bus pateiktas kristalografinių parametrų išgavimo komponento kūrimas.

7.1. Parametrų išgavimo komponentas

Parametrų išgavimo komponentas yra realizuotas pasiūlytu klausimų-atsakymų modelio pagrindu. Pasirinkimą lėmė potenciali galimybė išplėsti apmokymo duomenų rinkinį kokybiškais bendros tematikos klausimynais. Taip pat klausimų-atsakymų modelis siūlo papildomą lankstumą – parametrai, kurių pavyzdžių yra mažiau (arba nėra) duomenų rinkinyje galimai bus identifikuoti ir išgauti.

7.1.1. Klausimų-atsakymų duomenų rinkinio ruošimas

Kristalografinių tekstų buvo ieškoma arXiv talpykloje. Pateikus *"crystal structure" "space group"* paieškos užklausą buvo išgauti 164-ių straipsnių pirminiai LaTex dokumentai. Iš 164-ių straipsnių aibės "htlatex" įrankio pagalba pavyko nuskaityti 116 dokumentų. Nuskaityti straipsniai buvo peržiūrėti paruoštu pagalbiniu "corpus-editor" įrankiu. Įrankio pagrindas – serveris, kuriam yra pateikiamas JSON dokumentas talpinantis tekstų paketą. Naudotojas pildo dokumentą klausimų-atsakymų poromis per grafinę naudotojo sąsają naršyklėje ir galutinis duomenų rinkinys yra pateikiamas atsisiuntimui. Įrankio pagalba galima atrinkti tinkamus tekstus, formuoti klausimus tekstui ir žymėti teksto diapazonus. Įrankis taip pat vizualiai pažymi 512 morfemų ribą taikant BERTbase modelio žodyną. Pradžioje naudotojas pasirenka arba pats pažymi cheminį junginį, apie kurį norėtų formuoti klausimus. Vėliau įrankis pateikia junginių parametrų klausimus. Į klausimus naudotojas atsako žymėdamas žymekliu teksto diapazonus. Tokiu būdu atsakymo simbolių seka ir teksto diapazono pradžios indeksas yra įrašomi į JSON formato dokumentą.

Tokens 498
Edited false
url https://journals.iucr.org/b/issues/2022/04/00/tq5003/index.html
sources
pdf https://journals.iucr.org/b/issues/2022/04/00/tq5003/tq5003.pdf
doi https://doi.org/10.1107/S2052520622006254

Role of lone-pair electron localization in temperature-induced phase transitions in mimetite

The crystal structure of mimetite Pb5(AsO4)3Cl, a phosphate with apatite structure-type has been investigated in situ at 123, 173, 273, 288, 353 and 393 K. A careful inspection of the diffraction pattern and subsequent structure refinements indicated that mimetite transforms from the monoclinic to the hexagonal polymorph with increasing temperature. At 123 K, a monoclinic superstructure, mimetite-2M, with cell parameters a = 20.4487 (9), b = 7.4362 (2), c = 20.4513 (9) Å, β = 119.953 (6)°, V = 2694.5 (2) Å3 and space group P21 was observed. From 173 to 353 K, the reflections of the supercell were evident only along one direction of the corresponding hexagonal apatite-cell and the structure transforms to the polymorph mimetite-M with space group P21/b and unit-cell parameters a = 10.2378 (3), b = 20.4573 (7), c = 7.4457 (2) Å, β = 120.039 (5)°, V = 1349.96 (9) Å3. Only at higher temperature, i.e. 393 K, does mimetite adopt the hexagonal space group P63/m characteristic of apatite structure-types. The role of the electron lone pairs of Pb atoms in the phase transition was investigated through the analysis of the electron localization function (ELF) calculated based on the DFT-geometry optimized structures of the three polymorphs. The changes in spatial distribution of the 6s2 electron density during the phase transitions were explored by means of the Wannier Function Centres (WFCs) derived from ab initio molecular dynamics trajectories. In the high-temperature hexagonal structure the 6s2 electrons are spherically symmetric relative to the position of Pb atoms. At low temperature the maximum of 6s2 electron density is displaced relative to the position of Pb atom contributing to the polar interaction in the monoclinic polymorphs.
Models did not recognize chemical entity? Freeze
Submit
unclear_context transition needs_title unknown_chars valuable_case b
mimetite roberta-bc5cdr
mimetite Pb5(AsO4)3Cl roberta-bc5cdr space group mimetite?
phosphate roberta-bc5cdr 757 P21/b
phosphates roberta-bc5cdr
A5(TO4)3X roberta-bc5cdr Freeze Remove Add title in front

8 pav. ”corpus-editor” naudotojo grafinė sąsaja

Išgauti straipsniai pasirodė netinkami. Dauguma išgautų tekstų dėstė junginių kristalografinių parametrų kaitą skirtingomis sąlygomis, kintant spaudimui arba temperatūrai. Tokie tekstai kelia papildomus iššūkius, kadangi informacijos išgavimo sekoje taip pat numato temperatūros ir spaudimo reikšmes. Tiksliniai tekstai tiria junginių savybes įprastomis sąlygomis prie 293K temperatūros ir 100 kPa spaudimo sąlygų. Tikslinių tekstų reikalavimus geriau atitiko Acta Crystallographica E žurnale publikuojami tyrimai. Dauguma straipsnių šiame žurnale buvo pasiekiami be apribojimų ir kartu su HTML formato dokumentais. Per 7 dienas buvo nuskaitytas 20821 straipsnis (3,57GB duomenų). Taip pat tinkamais dokumentais pasirodė straipsnių santraukos prieinamos ACS Publications talpykloje. Iš jos pavyko nukelti 6607 straipsnių santraukų, publikuotų ”Inorganic Chemistry”, ”Journal of the American Chemical Society”, ”Organometallics”, ”Crystal Growth & Design” ir kituose kristalografiniuose žurnaluose. Išgauti kristalografiniai tekstai buvo paversti pastraipų masyvais. Viso pavyko išgauti 138236 kristalografinio teksto pastraipas.

Tokį kiekį duomenų rankiniu būdu apdoroti pasirodė neracionalu, todėl buvo nuspręsta ieškoti būdų žymėti tekstų diapazonus Crystallography Open Database pagalba. Duomenų bazėje jau figūravo junginių parametrai susieti su dalimi nuskaitytų straipsnių.

Crystallography Open Database duomenų bazėje pagrindinė lentelė ”data” talpina 496255 įrašus apie junginius. Junginiai lentelėje kartojasi, kadangi įrašas bazėje atitinka vieną nuskaityta dokumentą (mokslinį straipsnį, CIF formato dokumentą arba kitą šaltinį). Kiekvienas įrašas gali turėti 70 atributų, tarp kurių yra ir parametrai, kuriuos bus bandoma nuskaityti darbo metu:

”a”, ”b”, ”c” stulpeliai. Dešimtainės trupmenos nurodančios atitinkamai ”a”, ”b” ir ”c” gardelės konstantų dydžius angstremais (angstremas atitinka 0,1 nanometrą).

”siga”, ”sigb”, ”sigc” stulpeliai. Dešimtainės trupmenos nurodančios atitinkamai ”a”, ”b” ir ”c” gardelės konstantų reikšmių paklaidą. Pavyzdžiui, tekste gardelės parametro reikšmės paskutinio skaitmens po kablelio paklaida dažnai yra nurodoma skaičiumi skliaustuose, pavyzdžiui ”a=7.8783 (2) Å”. Šis užrašymas atitinka $a = 7{,}8783 \pm 0{,}0002$ (Å). ”data” lentelėje ”siga” stulpelis pateiktą pavyzdį talpintų $\{a = 7.8783, siga = 0.0002\}$.

”alpha”, ”beta”, ”gamma” stulpeliai. Dešimtainės trupmenos nurodančios atitinkamai ”α”, ”β” ir ”γ” gardelės kampus laipsniais.

”sigalpha”, ”sigbeta”, ”siggamma” stulpeliai. Dešimtainės trupmenos nurodančios atitinkamai ”α”, ”β” ir ”γ” gardelės kampų reikšmių paklaidą.

”sg”, ”sgHall”, ”sgNumber” stulpeliai. Šie stulpeliai nurodo įraše minimo cheminio junginio simetrijos grupę. ”sgNumber” skaičius yra pirminis raktas ”spacegroups” lentelėje, kuri talpina skirtingus atitinkamos grupės užrašymo būdus (Hall, Schoenflies, Hermann-Mauguin notacijos) ir žodinę simetrijos klasifikaciją. ”sg” stulpelis ”data” lentelėje nurodo būtent Hermann-Mauguin formato simetrijos grupę.

Junginį nurodantys stulpeliai. ”data” lentelėje cheminius junginius identifikuoja ”commonname” (junginio pavadinimas), ”mineral” (mineralo pavadinimas), ”formula” (cheminė formulė) stulpelių reikšmes. Kadangi cheminius junginius galima nurodyti daugeliu skirtingu būdų, straipsnyje sutikti cheminio junginio paminėjimai gali neatitikti duomenų bazėje laikomas reikšmes.

Klausimų-atsakymų aibė buvo formuojama tokiu būdu: tarp visų išgautų kristalografinio teksto pastraipų buvo ieškomos pastraipos, kurios turėtų paminėtą junginį ir tekste būtų parametro reikšmės atitikimas su esančia duomenų bazėje. Iš 138236 anksčiau nuskaitytų pastraipų tik 420 pastraipų atitiko tokią sąlygą. Atrinktų pastraipų pagrindu buvo suformuotos 1215 sekos.

Bandymai parodė, kad BERTbase modelis apmokytas SQuAD2.0 duomenų rinkiniu geriausiai pasirodė formuojant klausimus tokiu būdu:

- “What is the space group of «junginio_pavadinimias»?”
- “What is the a parameter of «junginio_pavadinimias»?”
- “What is the b parameter of «junginio_pavadinimias»?”
- “What is the c parameter of «junginio_pavadinimias»?”

7.1.1.1. Dirbtinis duomenų rinkinio plėtimas

Duomenų praplėtimu (angl. data augmentation) vadinamas apmokymo aibės plėtimas be naujų duomenų įtraukimo. Šie veiksmai yra taikomi, kai tinkamų duomenų kiekis yra nedidelis ar-

ba nėra prieinamas. Tokios technikos yra skirstomos į pagrįstas taisyklėmis, interpoliacijomis ir modeliais [FGW+21]. Aprašomai aibei buvo pritaikyta taisyklių technika dokumento (pastraipos) lygyje. Kadangi buvo pastebėta, kad pastraipose ieškomi cheminiai buvo sutinkami retai, buvo nuspręsta kiekvienos pastraipos pradžioje pridėti straipsnio, kuriam priklauso pastraipa, pavadinimą. Kristalografinių tyrimų moksliniuose straipsniuose, tiriamasis cheminis junginys dažnai yra paminimas straipsnio pavadinime. Tokiu būdu, pastraipų kiekis, kuriose yra minimi junginys ir parametro reikšmė, išaugo beveik dvigubai (nuo 420 iki 739 pastraipų). Pavyko suformuoti 1600 pastraipos-klausimo-atsakymo sekas. Gautas duomenų rinkinys toliau bus vadinamas CQAD (angl. Crystallography Question Answering Dataset).

Aprašytas duomenų rinkinio plėtimas taip pat turėjo teigiamą šalutinį poveikį – pastraipos, kuriose nurodoma į junginį pavadinime (angl. the title compound) modeliui būtų nesusiejamos su junginiu. Kadangi BERT modelio morfemų vektorizavimo procesas taip pat įdeda (angl. embeds) ir morfemos poziciją, modelis gebės suformuoti ryšį tarp ”the title compound” žodžių ir pirmame pastraipos sakinyje esančio junginio pavadinimo. Tokie duomenų rinkinio pakeitimai įtakos pirminį duomenų apdorojimą (angl. data pre-processing).

7.1.1.2. Neigiami pavyzdžiai

Kitas svarbus žingsnis yra papildyti CQAD tokiais klausimais, kurie neturėtų teisingo atsakymo pateiktame kontekste, arba ”neigiamais pavyzdžiais”. [RJL18] atskiria 6 skirtingas priežastis, dėl kurių klausimai gali neturėti atsakymo pateiktame kontekste:

- Neigiantis žodis yra įterptas į sakinį (pvz. niekada, neatitiko).
- Antonimai panaudojami klausime ir atsakyme.
- Skirtingos esybės minimos klausime ir pateiktame kontekste.
- Tarpusavyje nesuderinami klausimo ir atsakymo kontekstai. Pavyzdžiui, ”Įmonė pakėlė produktų A ir B kainą 10%. Klausimas: Kuris įmonės produktas yra dalinamas nemokamai?”
- Klausimas turi sąlygą, kurios netenkina jokia informacija iš konteksto.
- Neutrali priežastis. Kontekste nėra pateikto atsakymo. Pavyzdžiui, ”Junginio simetrijos grupė yra P1. Klausimas: Kokia yra junginio lydymosi temperatūra?”.

Akademiniuose tekstuose dalis aukščiau išvardintų atsakymo nebuvimo priežasčių yra mažai tikėtinos. Neigimo žodis yra retas reiškinys kristalografinių junginio parametrų kontekste. Surinktose pastraipose nebuvo rastas atvejis, kuriame būtų rašoma, kad “junginio struktūra nėra «simetrijos_grupė»”, todėl pildyti duomenų rinkinį tokiais pavyzdžiais pasirodė neracionalu. Antonimai ir tarpusavyje nesuderinamo klausimo ir atsakymo kontekstai taip pat nėra pritaikomi užduočiai – šiuo metu paruoštiems klausimams ir atsakymams antonimų įtakos būti negali, nes klausiama yra skaitinių arba tekstinių reikšmių. Kitaip nei aukščiau apžvelgtos atsakymo nebuvimo priežastys, klausimo su neįmanoma sąlyga atvejai gali būti aktualūs sudėtingesniam informacijos išgavimui, pavyzdžiui, jeigu tektų išgauti junginio savybes prie tam tikros temperatūros arba spaudimo.

Tokiu būdu lieka dvi aktualios priežastys teisingo atsakymo nebuvimui: skirtingos esybės klausime ir kontekste, ir neutralios priežastys. Paprasčiausias būdas papildyti esamą aibę tokio tipo

klausimais yra taikant duomenis iš Crystallography Open Database surasti visų pastraipų sąraše:

- tekstus, kuriuose junginys yra minimas, bet parametro tekste nėra
- tekstus, kuriuose parametrai yra nurodomi kito junginio, nei klausiama
- tekstus, kuriuose nekalbama apie junginius ir jų parametrus

SQuAD2.0 duomenų rinkinyje ”neigiamų pavyzdžių” yra 35%. Kadangi, kristalografinės informacijos išgavimo užduočiai kritiškai svarbus yra tikslumas (geriau yra neišgauti nieko, nei išgauti neteisingą informaciją), ”neigiamų pavyzdžių” gali prireikti daugiau. Nors ir buvo apžvelgti pagrindiniai netikslių atsakymų atvejai, visgi sunku yra nuspėti kokio ”neigiamų pavyzdžių” santykio pakaktų. Tikėtina, kad pirmieji treniruoto modelio bandymai leis suprasti, kokiais atvejais modelis klysta ir remiantis klaidomis bus koreguojamas ”neigiamų pavyzdžių” kiekis ir įvairovė.

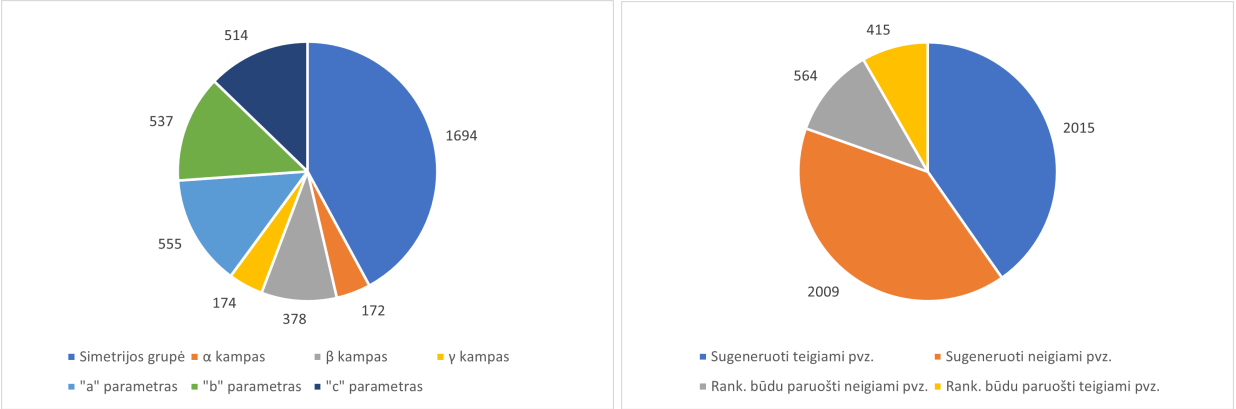
Bandymų eigoje buvo nustatytas automatiškai sugeneruotų neigiamų pavyzdžių trūkumas – neigiami pavyzdžiai nedengė atvejo, kai šalutinis junginys yra siejamas su parametrais tekste. Pavyzdžiu, (9) santraukos ištraukoje yra minimi junginiai $(NBu_4)_2[B_{12}H_{12}]$ ir $Li_2[B_{10}H_{10}]$, kurie buvo naudojami galutiniams junginiams gauti. Tokiu būdu, tekste minimi kristalografiniai parametrai yra siejami tik su galutiniu junginiu ($[Li(NH_3)_4]_2[B_{12}H_{12}]\cdot 2NH_3$).

The title compounds contain 20.04-52.23 wt % ammonia and 5.94-13.01 wt % hydrogen. They were synthesized in liquid ammonia, using $(NBu_4)_2[B_{12}H_{12}]$ and $Li_2[B_{10}H_{10}]$ as starting materials. $[Li(NH_3)_4]_2[B_{12}H_{12}]\cdot 2NH_3$ (1) crystallizes in the monoclinic crystal system (space group P21/c, $a = 9.183(2)$ Å, $b = 8.133(1)$ Å, $c = 16.375$ Å, $\beta = 110.54(1)°$, $V = 1143.97(40)$ Å^3, $Z = 2$).

9 pav. Minimi keli junginiai, bet parametrai siejami tik su vienu iš jų. Šaltinis: [KPM+13]

Šią problemą galima spręsti pridedant neigiamus apmokymo atvejus, kuriuose klausimai sieja šalutinius junginius su parametrais. Akivaizdžiausias būdas sugeneruoti tokius atvejus automatiškai yra pirma nustatyti visus cheminius junginius pateiktame tekste ir atmetus junginius figūruojančius COD, suformuoti likusiems junginiams neigiamus klausimų pavyzdžius su visais paieškos parametrais (gardelės matmenys, kampai, simetrinė grupė). Tokiu būdų sugeneruotų klausimų kokybė priklausytų nuo poros veiksnių: junginių žymėjimo tikslumo ir COD duomenų bazėje minimų junginių atitikimu su tekste esančiais junginiais.

Kadangi darbo eigoje buvo pastebėta, jog dažnas cheminis junginys tekste tiksliai neatitinka junginių COD duomenų bazėje (pvz. dėl dalį žyminčių formulių) buvo nuspręsta duomenų rinkinį papildyti rankiniu būdu paruoštais atitinkamais neigiamais klausimais. Šiam tikslui buvo pasirinktos 25 pastraipos, kurios leido paruošti 564 ”neigiamus” klausimus ir 415 klausimus su atsakymais.



10 pav. CQAD_50 duomenų rinkinio sudėtis

7.1.1.3. Kitos duomenų praplėtimo galimybės

CQAD duomenų rinkinys gali būti plečiamas taikant ir kitas taisykles. Taisykles galima suskirstyti lygiais: morfemos, sakinio ir pastraipos lygio. Morfemų lygio taisyklės galėtų pakeisti parametrų reikšmes. Nors tolimesni tyrimai parodė, kad modelis nesuformuoja stiprių ryšių su konkrečiomis reikšmėmis, šis duomenų papildymas galėtų turėti teigiamą įtaką galutiniam rezultatui. Akivaizdžių sakinio lygio papildymo taisyklių yra mažiau. Kadangi BERT ir kiti kontekstiniai modeliai vektorizuoja morfemų pozicijas, paprastas žodžių sumaišymas sakinio lygyje galimai turės žalingą poveikį modelio suformuotiems ryšiams. Tikslingiau būtų atlikti parametrų keitimą tarpusavyje, nors daugelyje straipsnių parametrų išdėstymo rikiavimas būna vienodas. Pastraipos lygio duomenų praplėtimas apimtų sakinių eilės keitimą, sakinių išėmimą ir pridėjimą. Kristalografinių tekstų apžvalga parodė, kad junginiai ir jų parametrai gali būti išdėstyti skirtingose pastraipos dalyse kartu su nuorodomis. Šios sąlygos apsunkina pastraipos lygio duomenų generavimą, kadangi nuorodos ir junginio ryšiai bei kertinių elementų eilė turi išlikti nepakitę.

Aukščiau išvardintos duomenų praplėtimo taisyklės pritaikytos nebuvo.

7.1.1.4. Klausimų-atsakymų modelių apmokymas

BERT architektūra paremto diapazonų paieškos tikslumą, gali įtakoti du pagrindiniai veiksniai:

- Kalbos modelio artumas tikslinei kalbai.
- Klausimų-atsakymų apmokymo duomenų rinkinių įvairovė: klausimų tematika ir ”neigiamų pavyzdžių” kiekis bei kokybė.

Remiantis įvardintais veiksniais buvo suformuoti modeliai, kurių tikslumas leis suprasti veiksnių įtaką rezultatams. Kalbos modelių skirtumas bus realizuotas BERTbase, BioBERT ir SciBERT modelių pagalba. Visi naudojami modeliai skiria didžiąsias raides nuo mažųjų (angl. cased model), kadangi kristalografiniuose tekstuose simbolių registras talpina svarbia kontekstinę informaciją. Taip pat išvardinti modeliai žodynuose talpina simbolius, kurie yra dažnai sutinkami

tiksliniuose tekstuose (pvz”Å”, ”°”, ”α”, ”β”, ”γ”). Klausimų-atsakymų duomenų rinkinių įvairovė bus įgyvendinta pateikiant modeliams kelias duomenų rinkinių variacijas. Vienais atvejais BERTbase, BioBERT ir SciBERT modeliai bus apmokomi tik CQAD_50 duomenų rinkiniu, turinčiu 50% neigiamų pavyzdžių, kitais – BERTbase, BioBERT ir SciBERT modeliai bus apmokyti SQuAD2.0 duomenų rinkiniu kartu su CQAD_50. Tokiu būdu gautos variacijos yra:

BERTbase + CQAD_50. BERTbase kalbos modelis apmokytas CQAD_50 duomenų rinkiniu.

BERTbase + SQUAD2.0 + CQAD_50. BERTbase kalbos modelis apmokytas SQUAD2.0 ir CQAD_50 duomenų rinkiniais.

BioBERT + CQAD_50. BioBERT kalbos modelis apmokytas CQAD_50 duomenų rinkiniu.

BioBERT + SQUAD2.0 + CQAD_50. BioBERT kalbos modelis apmokytas SQUAD2.0 ir CQAD_50 duomenų rinkiniais.

SciBERT + CQAD_50. SciBERT kalbos modelis apmokytas CQAD_50 duomenų rinkiniu.

SciBERT + SQUAD2.0 + CQAD_50. SciBERT kalbos modelis apmokytas SQUAD2.0 ir CQAD_50 duomenų rinkiniais.

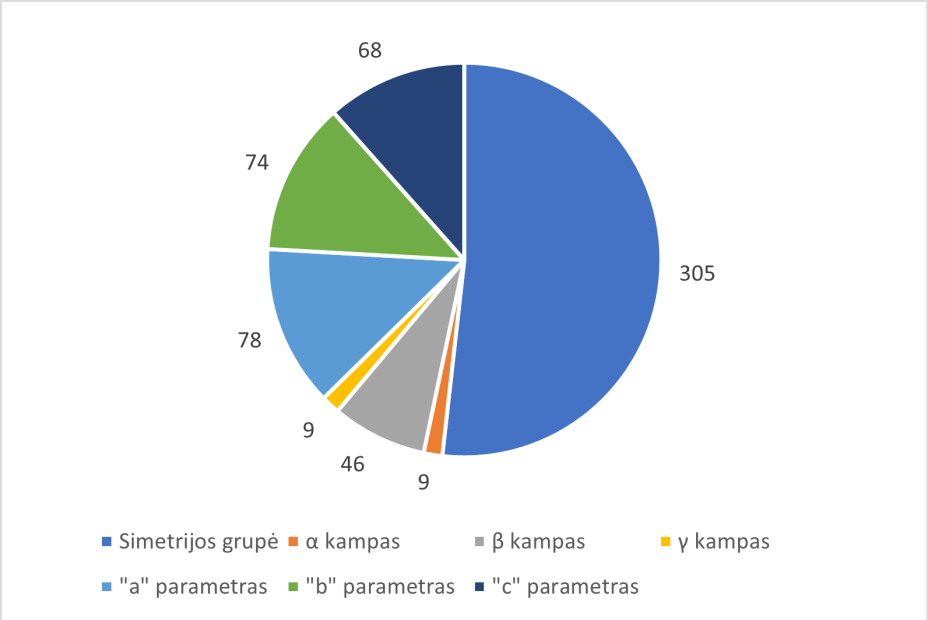
Visi modeliai buvo apmokyti CQAD_50 taikant vienodus hiperparametrus: 3 epochos, paketų dydis (angl. batch size) lygus 3, mokymosi dažnis (angl. learning rate) lygus 0,00005, modelio apšilimas (angl. warmup) lygus 0. SciBERT + SQUAD2 modelis, kitaip nei kiti SQUAD2 apmokyti modeliai, nėra laisvai prieinamas https://huggingface.co modelių talpykloje, todėl SciBERT + SQUAD2 variacija buvo apmokyta šio darbo ribose (apmokymas vyko 2 val. 36 min. Google Colab aplinkoje naudojant Tesla T4 vaizdo plokštę).

Modelis	Pavadinimas	Šaltinis
BERTbase	bert-base-cased	https://huggingface.co/bert-base-cased
BERTbase + SQUAD2.0	deepset/bert-base-cased-squad2	https://huggingface.co/deepset/bert-base-cased-squad2
BioBERT	monologg/biobert_v1.1_pubmed	https://huggingface.co/monologg/biobert_v1.1_pubmed
BioBERT + SQUAD2.0	ktrapeznikov/biobert_v1.1_pubmed_squad_v2	https://huggingface.co/ktrapeznikov/biobert_v1.1_pubmed_squad_v2
SciBERT	allenai/scibert_scivocab_cased	https://huggingface.co/allenai/scibert_scivocab_cased

2 lentelė. Bandymams naudotos modelių realizacijos

7.1.1.5. Gautų modelių vertinimas

CQAD_50 buvo atsitiktinai padalinta į du poaibius: 20% duomenų rinkinio buvo skirta modelio tikrinimui ir likę 80% modelio apmokymui. Tikrinimo poaibyje viso yra 589 klausimų iš kurių 280 yra klausimai, neturintys atsakymo



11 pav. CQAD_50 tikrinimo aibės sudėtis.

Pradiniai rodikliai

Pradžioje buvo įvertinti BERTbase + SQUAD2, BioBERT + SQUAD2 ir SciBERT + SQUAD2 modeliai. Prasčiausiai pasirodė bazinis kalbos modelis BERTbase + SQUAD2. Modelio faktinis tikslumas siekė vos 18,87. BioBERT ir SciBERT apmokytų bendrais klausimais pasirodė panašiai – nors faktinis tikslumas yra lygus 30,77 ir 35,06 atitinkamai, teigiamų pavyzdžių atitikimų daugiau yra BioBERT + SQUAD2 atveju (30,74% prieš 36,47%). Neigiamus pavyzdžius geriausiai identifikavo SciBERT + SQUAD2 modelis – 81,79%.

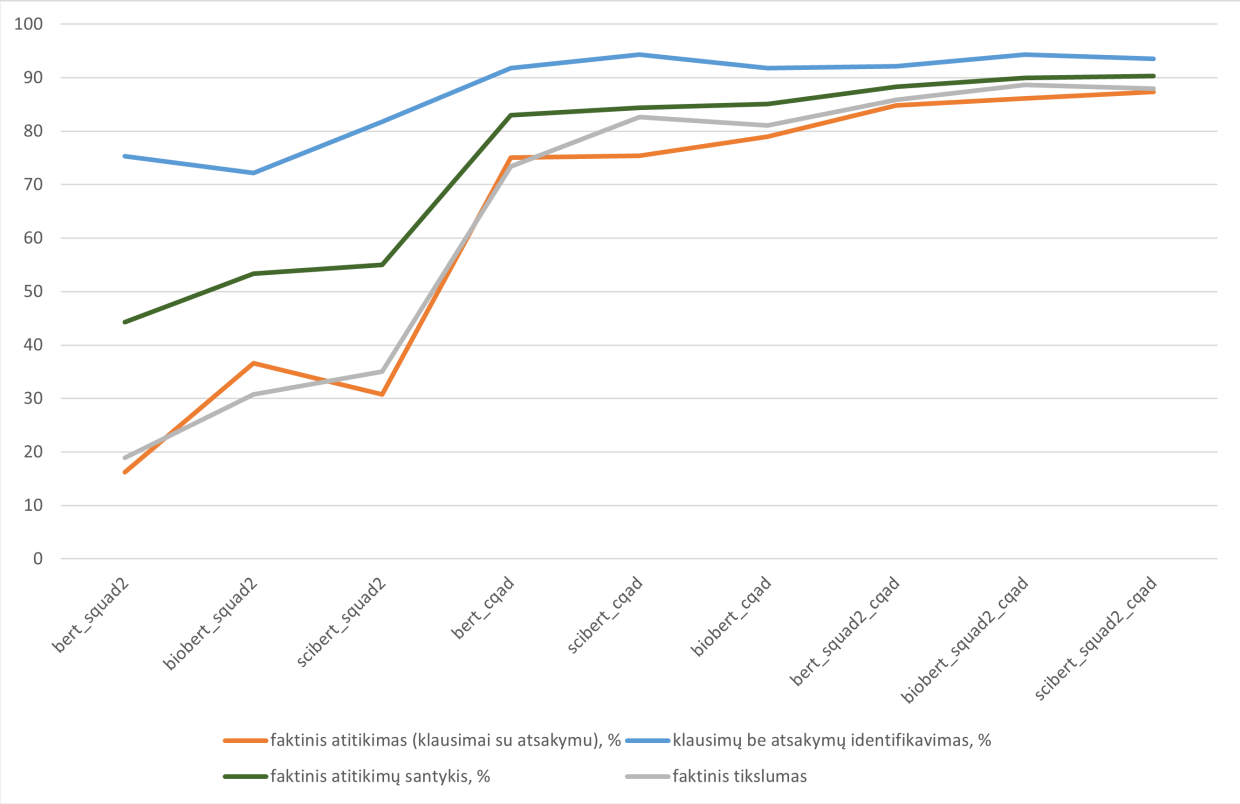
CQAD rezultatai

BERTbase, BioBERT ir SciBERT modeliai pritaikyti klausimų-atsakymų užduočiai ir apmokyti CQAD duomenų rinkiniu parodė rezultatus lenkiančius pradinius rodiklius: faktinis atsakymų tikslumas išaugo dvigubai visais kalbos modelių atvejais ir didžiausia tikslumo reikšmę parodė SciBERT + CQAD modelis (82,62%). Neigiamų pavyzdžių identifikavimas taip pat geriausiai sekėsi SciBERT + CQAD modeliui (94,29%). Tuo tarpu atitikimų santykis liko geriausias BioBERT + CQAD atveju (85,06%). Šio bandymo metu kalbos modelio įtaka pasirodė menkesnė.

SQUAD2.0 + CQAD rezultatai

Šiam bandymui buvo naudojami modeliai apmokyti 100K+ bendro pobūdžio klausimų (SQuAD2.0) ir 2456 kristalografiniais klausimais (CQAD).

Kalbos modeliai pritaikyti akademiniams tekstams parodė geresnius rodiklius visais tiriamais rodikliais. BioBERT + SQUAD2.0 + CQAD ir SciBERT + SQUAD2.0 + CQAD modelių faktinis tikslumas siekė 88,67 ir 87,95 atitinkamai. Bendrinio kalbos modelio rezultatas buvo kiek žemesnis – 85,90. Neigiamus klausimų pavyzdžius geriausiai identifikavo BioBERT modelio variacija (94,29%), tuo tarpu tikslių atitikimų daugiau parodė SciBERT modelis (90,32%).



12 pav. Modelių vertinimų grafikas

modelis	teig. pvz. atitik., %	n. pvz. iden-imas, %	atitik., %	lingv. F1	fakt. P
bert+squad2	16,18	75,36	44,31	47,57	18,87
biobert+squad2	36,57	72,14	53,31	61,98	30,77
scibert+squad2	30,74	81,79	55,01	60,01	35,06
bert+cqad	75,08	91,79	83,02	83,02	73,42
scibert+cqad	75,40	**94,29**	84,38	84,38	82,62
biobert+cqad	78,96	91,79	85,06	85,13	81,06
bert+squad2+cqad	84,79	92,14	88,29	88,29	85,90
biobert+squad2+cqad	86,08	**94,29**	89,98	90,05	**88,67**
scibert+squad2+cqad	**87,38**	93,57	**90,32**	**90,32**	87,95

3 lentelė. Bandymų rezultatai

Rezultatų analizė

SQuAD2.0 bendriniais klausimais apmokytas modelis geba teisingai išgauti parametrus 36% atvejų. Pirminis bandymas parodė kalbos modelio pritaikymo vertę – modelis, kuris yra pritaikytas (angl. fine-tuned) mokslinei literatūrai ir apmokytas bendro pobūdžio klausimais, daugeliu rodiklių atžvilgiu veikia tiksliau, nei pirminis kalbos modelis, tačiau pritaikytas modelis vietomis dažniau klysta tais atvejais, kai atsakymo tekste nėra. Tai siejama su tekstais, naudotais BioBERT apmokymui: akademiniai PubMed tekstai rašo apie ”simetrija” ir ”grupes” daugelyje skirtingų kontekstų. PubMed grąžina virš 350 tūkstančių straipsnių turinčiu santraukoje žodį ”space” ir 3 milijonus straipsnių mininčių žodžiu ”group”. Tokiu būdu, BioBERT modelis labiau sieja morfemas su šiais

žodžiais, nei modelis apmokytas tik knygomis bendromis temomis ir Wikipedia straipsniais ir todėl yra gaunama daugiau neteisingai teigiamų išvesčių ”neigiamų pavyzdžių” atvejais.

Visus pirminius rodiklius pagerino CQAD apmokyti modeliai. Kalbos modelio įtaka šiame bandyme sumažėjo – BioBERT, BERTbase ir SciBERT modeliai parodė panašius rezultatus, bet moksliniams tekstams pritaikyti modeliai vis gi pasirodė geriau. Skirtumai tarp kalbos modelių rezultatų dar labiau sumažėjo tikrinant SQUAD2.0 + CQAD apmokytus modelius. Visų kalbos modelių SQUAD2.0 + CQAD variacijos parodė geresnius rezultatus, nei CQAD. Dideliu kiekiu (daugiau 100 tūkstančių) bendrinių klausimų ir CQAD apmokytas modelis geriau žymėjo atsakymų diapazonus sudėtiniuose sakiniuose, kurių pavyzdžių yra SQuAD2.0 duomenų rinkinyje. Taip pat rodikliai SQUAD2.0 + CQAD yra geresni dėl sąlyginai nedidelio kiekio klausimų CQAD susijusiu su gardelės kampais. SQuAD2.0 pavyzdžiai tokiu būdu kompensuoja gardelės matmenų pavyzdžių stoką.

A series of ternary Zintl phases, Ca2CdP2, Ca2CdAs2, Sr2CdAs2, Ba2CdAs2, and Eu2CdAs2, have been synthesized through high temperature metal flux reactions, and their structures have been characterized by single-crystal X-ray diffraction. They belong to the Yb2CdSb2 structure type and crystallize in the orthorhombic space group Cmc21 (No. 36, Z = 4) with cell dimensions of a = 4.2066(5), 4.3163(5), 4.4459(7), 4.5922(5), 4.4418(9) Å; b = 16.120(2), 16.5063(19), 16.904(3), 17.4047(18), 16.847(4) Å; c = 7.0639(9), 7.1418(8), 7.5885(11), 8.0526(8), 7.4985(16) Å for Ca2CdP2 (R1 = 0.0152, wR2 = 0.0278), Ca2CdAs2 (R1 = 0.0165, wR2 = 0.0290), Sr2CdAs2 (R1 = 0.0238, wR2 = 0.0404), Ba2CdAs2 (R1 = 0.0184, wR2 = 0.0361), and Eu2CdAs2 (R1 = 0.0203, wR2 = 0.0404), respectively. Among these, Ca2CdAs2 was found to form with another closely related structure, depending on the experimental conditions—monoclinic space group Cm (No. 8, Z = 10) with lattice constants a = 21.5152(3) Å, b = 4.30050(10) Å, c = 14.3761(2) Å and β = 110.0170(10)° (R1 = 0.0461, wR2 = 0.0747).

13 pav. CQAD_50 vertinimo aibės konteksto pavyzdys. Šaltinis: [WYP+11]

Iš konteksto (Pav. 13) ”c” gardelės konstantą Ca_2CdP_2 junginiui pavyko pažymėti tik SQUAD2.0 apmokytiems modeliams (SciBERT + SQUAD2 + CQAD, BioBERT + SQUAD2 + CQAD, BioBERT + SQUAD2). Tai rodo, kad bendriniai SQUAD2 klausimai leido modeliui tiksliau nuskaityti kontekstus, kuriuose yra nurodoma į tam tikrą junginį su numeriu. Likę modelių žymėjimai klaidingai nurodė į ”14.3761(2)” reikšmę. Taigi, SQuAD2.0 duomenų rinkinys kompensuoja CQAD duomenų rinkinio trūkumus.

Išvados

Bandymų metu buvo panaudoti skirtingų variacijų modeliai ir jų pritaikymo rezultatai leidžia geriau suprasti skirtingų faktorių įtaką galutiniams rezultatams. Gauti rezultatai byloja, kad 50% ”neigiamų pavyzdžių” apmokyme yra pakankamas kiekis rodantis tenkinančius rezultatus. Taip pat bandymai įrodo ir kalbos modelio adaptacijos naudą tikslumui. Papildomas pastaba apie apmokymų aibes: apmokymo duomenų rinkinio klausimų įvairovė ir dydis taip pat teigiamai įta-

koja rezultatus, o bendrinių duomenų rinkinių įtraukimas gali ženkliai pagerinti modelio galimybės nuskaitant sudėtingas sakinių konstrukcijas.

Galutiniame darbo sraute bus naudojamos dvi geriausia tikslumą parodžiusios modelio realizacijos: BioBERT + SQUAD2 + CQAD ir SciBERT + SQUAD2 + CQAD.

7.1.2. Cheminių junginių išgavimo komponento realizacija

Cheminių junginių išgavimo komponentas taip pat bus pagrįstas TNN modeliais. Dažnai junginį nuo kitų teksto sekų padeda atskirti kontekstas, todėl reguliariosiomis išraiškomis paremti sprendimai neužtikrins pakankamo tikslumo.

7.1.2.1. Duomenų rinkinio formavimas

Duomenų rinkinys šiam komponentui turi atitikti cheminių junginių užrašymo ypatumus ir jų variacijas tiksliniuose tekstuose: įprastas chemines formules, IUPAC nomenklatūrą, liekanas žyminčias chemines formules ir apibendrinančias formules.

Įprastų cheminių formulių žymėtus pavyzdžius buvo nuspręsta paruošti remiantis BC4CHEMD [KRL+15] duomenų rinkiniu. Rinkinys talpina 10000 PubMed straipsnių santraukų, kuriose yra pažymėti 84355 junginiai. Anotacijos taip pat yra klasifikuojamos: santrumpos (angl. abbreviations), junginių šeimos, formulės (molekulinės, SMILES, InChI), unikalūs identifikatoriai, IUPAC ir trivialūs pavadinimai turi specialius žymėjimus. Tai leidžia nesunkiai atrinkti konkrečiai užduočiai svarbius pavyzdžius. Atmetus unikalius identifikatorius (identifikatoriaus unikalumas gali būti apibrėžtas vienos duomenų bazės/žurnalo ribose, todėl bendrai junginių duomenų bazei tokie identifikatoriai nėra tinkami) ir junginių šeimas (apibendrinančios sąvokos) buvo atrinkta 10000 anotuotų santraukų. Taip pat duomenų rinkinys buvo papildomai pakoreguotas, kad atitiktų tikslinį tekstą (pvz. apatinio registro skaičiai formulėse paversti įprastais). Antras panaudotas duomenų rinkinys [KKF+08] yra labiau orientuotas į IUPAC notacija užrašomus cheminius junginius. Rinkinys talpina 100 rankiniu būdu anotuotų MEDLINE straipsnių santraukų. Panašiai kaip ir BC4CHEMD atveju, unikalūs identifikatoriai ir junginių šeimos buvo pašalintos iš duomenų rinkinio apdorojimo metu.

Kitas žingsnis reikalavo papildyti duomenų rinkinį liekanas žyminčiomis cheminėmis formulėmis (pvz. [Fe(C_5H_5)($C_{30}H_{32}O_7P$)] [ŠLC09]) ir apibendrinančiomis formulėmis (pvz. $Na_2SeO_4 \cdot 1.5H_2O$ [For15]). Tokie cheminių formulių užrašymai talpina daugiau informacijos, nei molekulinės formulės, todėl dvipusis formulių konvertavimas nėra įmanomas. Kristalografinių tekstų pastraipos su minėtomis formulėmis buvo pažymėtos rankiniu būdu. Buvo anotuotos 23 pastraipos, kurios dėl skirtingų priežasčių nepateko į klausimų-atsakymų duomenų rinkinį. Pateikus nors ir nedidelį kiekį rankiniu būdu anotuotų kristalografinių tekstų, buvo pastebėtas tikslesnis veikimas keliai aspektais. Pavyzdžiui, modelis apmokytas ankstesniu duomenų rinkiniu klaidingai klasifikuodavo cheminio junginio žymę kaip formulės dalį. Papildytas duomenų rinkinys leidžia modeliui išvengti šios klaidos.

For [(NH3)5Ru(NC5H5)]Cl3·1.4H2O (2): orthorhombic space group Pnma, Z = 4, a = 22.667(12) Å, b = 7.095(2) Å, c = 10.097(8) Å.

14 pav. Netikslaus cheminio junginio žymėjimo pavyzdys. Modelis klaidingai įtraukia junginio žymę dėl panašių pavyzdžių stokos duomenų rinkinyje. Teksto šaltinis: [SSB+97]

Taigi, galutiniame cheminių junginių atpažinimo duomenų rinkinyje yra 10123 pastraipos (70127 junginiai). Junginių klasifikavimo buvo atsisakyta, todėl junginiai užrašyti skirtingomis notacijos yra pažymėti vienodai (”COMPOUND”). Toliau duomenų rinkinys bus vadinamas CCD (angl. Chemical Compound Dataset).

7.1.2.2. Cheminių junginių išgavimo modelių apmokymas

Cheminių junginių išgavimui taip pat buvo naudojamas BERT architektūros modelis pritaikytas morfemų klasifikavimo užduočiai. Modelio tikslas — apdoroti fiksuoto ilgio morfemų įvestį ir priskirti įvesties morfemoms klases ”B-COMPOUND” (junginio pradžios žymėjimas), ”I-COMPOUND” (junginio tęsinys) ir ”O” (likusios morfemos). Kaip ir klausimų-atsakymų užduočiai buvo pasirinkti BERTbase, BioBERT ir SciBERT kalbos modeliai.

BERTbase + CCD. BERTbase kalbos modelis apmokytas CCD duomenų rinkiniu.

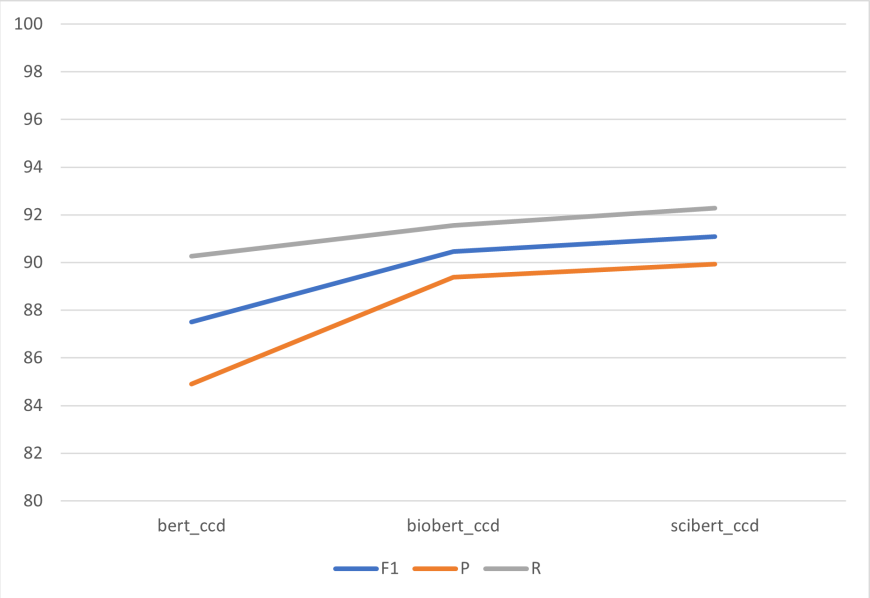
BioBERT + CCD. BioBERT kalbos modelis apmokytas CCD duomenų rinkiniu.

SciBERT + CCD. SciBERT kalbos modelis apmokytas CCD duomenų rinkiniu.

Visi jie palaiko 512 morfemų įvesti ir apmokyti tuo pačiu duomenų rinkiniu ir vienodais hiperparametrais: 3 epochos, paketų dydis (angl. batch size) lygus 3, mokymosi dažnis (angl. learning rate) lygus 0,00005, modelio apšilimas (angl. warmup) lygus 0.

Gautų modelių vertinimas

Modelių tikrinimui buvo skirta 20% gauto duomenų rinkinio. Dėl nedidelio kiekio anotuotų kristalografinių tekstų buvo nuspręsta juos pateikti tik apmokymui – tikrinimo aibėje bus tik BC4CHEMD [KRL+15] ir [KKF+08] duomenų rinkinių įrašai.



15 pav. Modelių vertinimų grafikas

Modelis	F1	Tikslumas	Atkūrimas
bert+ccd	87,5	84,9	90,3
biobert+ccd	90,5	89,4	91,6
scibert+ccd	**91,1**	**89,9**	**92,3**

4 lentelė. Bandymų rezultatai

Junginių paieškos komponento realizacija SciBERT kalbos modelio pagrindu parodė aukščiausius tikslumo, atkūrimo ir F1 statistikos rodiklius. Šie rodikliai nebūtinai bus vienodi taikant modelį kristalografiniams tekstams – vertinimo duomenų rinkinys apima biomedicininius straipsnių imtį, tačiau gauti rezultatai leidžia prognozuoti modelių kokybę tikslinėje tekstų sferoje. Kitame skyriuje SciBERT + CCD ir BioBERT + CCD modeliai bus bandomi galutiniame parametrų išgavimo konvejeryje.

7.1.2.3. Komponentų sujungimas

Taigi, galutinis parametrų išgavimo konvejeris yra sudarytas iš atskirų junginių ir parametrų išgavimo komponentų. Tyrimų metu išaiškėjo galimi konvejerio sekų kandidatai:

- SciBERT + CCD ir SciBERT + SQUAD2 + CQAD
- BioBERT + CCD ir BioBERT + SQUAD2 + CQAD
- SciBERT + CCD ir BioBERT + SQUAD2 + CQAD
- BioBERT + CCD ir SciBERT + SQUAD2 + CQAD

7.1.2.4. Konvejerių vertinimas

Automatinis duomenų rinkinio formavimo būdas puikiai tiko apmokymo aibės kūrimui — junginio ir parametrų ryšiai išgauti iš COD duomenų bazės buvo pakankamai tikslūs, bet dėl kelių

priežasčių toks duomenų rinkinys nėra tinkamas viso konvejerio vertinimui. Kadangi patikrintas bus viso sprendimo tikslumas ir atkūrimas, duomenų rinkinys turi talpinti visas galimas junginio-parametro poras sutinkamas kontekste. Automatiškai sugeneruotame duomenų rinkinyje:

- Yra fiksuojami tikslūs junginių sutapimai duomenų bazėje ir pateiktame kontekste. Tokiu būdu junginio užrašymas kita notacija nebus automatiškai pridėtas į duomenų rinkinį.
- Parametro reikšmė duomenų bazėje turi tiksliai atitikti reikšmę pateiktame kontekste. Jeigu straipsnio autorius nusprendė straipsnyje nurodyti parametrą mažesniu tikslumu, nei prisegtame CIF dokumente — junginio-parametro pora nebus identifikuota ir įtraukta į duomenų rinkinį.

Synthesis, Structure, and Reactivity of [RuCl(PP)L]PF_6 (PP = $(PPh_3)_2$, $Ph_2P(CH_2)_4PPh_2$; L = $P(py)_3$, $PPh(py)_2$, py = 2-pyridyl). The "Missing" *P,N,N′*-Coordination Mode for 2-Pyridylphosphines

Richard P. Schutte, Steven J. Rettig, Ajey M. Joshi, and Brian R. James*

Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, BC, Canada V6T 1Z1

Received July 3, 1997[⊗]

The complexes [RuCl(PPh_3)$_2$(*P,N,N′*-$PPh_{3-x}(py)_x$)]PF_6 (x = 2, **1b**; 3, **1c**; py = 2-pyridyl) were isolated from the reaction of $RuCl_2(PPh_3)_3$ with 1 equiv of $PPh_{3-x}(py)_x$ and NH_4PF_6 in acetone. Crystals of **1b** ($C_{52}H_{43}ClF_6N_2P_4$-Ru) are monoclinic, a = 17.795(2), b = 11.375(4), and c = 23.343(2) Å, β = 97.012(8)°, Z = 4, space group $P2_1/c$; those for **1c** ($C_{51}H_{42}ClF_6N_3P_4Ru$) are monoclinic, a = 17.812(1), b = 11.353(2), and c = 23.391(1) Å, β = 97.738(5)°, Z = 4, space gr… were refined by full-matrix lea… and 8121 reflections with $I \geq$ … previously unreported. The … $Ph_2P(CH_2)_4PPh_2$) were made by …

ic970835j_si_001.cif - Notepad

File Edit Format View Help

_cell_length_b 11.353(2)

_cell_length_c 23.3914(13)

16 pav. Reikšmės straipsnyje ir prisegtame CIF formato dokumente gali skirtis. Teksto šaltinis: [SRJ+97]

Dėl pateiktų priežasčių buvo nuspręsta suformuoti unikalų duomenų rinkinį rankiniu būdu. Corpus-editor įrankio pagalba buvo anotuoti 16 pastraipų, kurios nebuvo naudojamos konvejerio komponentų apmokymui ir vertinimui. Verta paminėti, kad tikrinimo duomenų rinkinys negali apimti visos parametrų dėstymo įvairovės dėl kuklaus pastraipų kiekio, tačiau dažniau sutinkamos dėstymo konstrukcijos vis gi reprezentuotos gautame duomenų rinkinyje. 16-oje pastraipų buvo pažymėtos visos (277) junginių-parametrų poros, kurių pagalba bus vertinamas viso konvejerio tikslumas.

7.1.2.5. Rezultatai

Konvejerių kandidatų palyginimui buvo pasirinktas junginio identifikavimo tikimybės slenkstis 0,9 ir parametro identifikavimo tikimybės slenkstis 0,9. Pirminis lyginimas vyks atsižvelgiant į F1 statistika ir geriausius rodiklius parodžiusiam konvejeriui slenksčiai bus pakoreguoti.

Konvejeris	Viso parametrų	Viso spėjimų	Neteisingi	Teisingi	Atkūrimas	Tikslumas	F1
SciBERT(NER) SciBERT(QA)	267	200	29	171	64,0	85,5	73,2
SciBERT(NER) BioBERT(QA)	267	185	**15**	170	63,7	**91,9**	75,2
BioBERT(NER) BioBERT(QA)	267	190	17	**173**	**64,8**	91,1	**75,7**
BioBERT(NER) SciBERT(QA)	267	203	30	**173**	**64,8**	85,2	73,6

5 lentelė. Bandymų rezultatai

Bandymai parodė, kad taikant 0,9 tikimybės slenkstį konvejerių komponentams, aukščiausius F1 statistikos rodiklius siekė BioBERT + SQUAD2 + CQAD paremtas sprendimas (SciBERT + CCD ir BioBERT + SQUAD2 + CQAD 0.752 ir BioBERT + CCD ir BioBERT + SQUAD2 + CQAD 0.757). Konvejeris patenkinamai išgavo didžiąją dalį parametrų iš kontekstų, kuriuose nuosekliai yra dėstomi junginiai ir atitinkami kristalografiniai parametrai.

> Hydrothermal Chemistry, Structures, and Luminescence Studies of Alkali Hafnium Fluorides. This paper describes the hydrothermal chemistry of alkali hafnium fluorides, including the synthesis and structural characterization of five new alkali hafnium fluorides. Two ternary alkali hafnium fluorides are described: Li2HfF6 in space group P31m with a = 4.9748(7) Å and c = 4.6449(9) Å and Na5Hf2F13 in space group C2/m with a = 11.627(2) Å, b = 5.5159(11) Å, and c = 8.4317(17) Å. Three new alkali hafnium oxyfluorides are also described: two fluoroelpasolites, K3HfOF5 and (NH4)3HfOF5, in space group Fm3m with a = 8.9766(10) and 9.4144(11) Å, respectively, and K2Hf3OF12 in space group R3m with a = 7.6486(11) Å and c = 28.802(6) Å.

17 pav. Parametrų išgavimas iš taikant SciBERT + CCD ir BioBERT + SQUAD2 + CQAD konvejerį ir 0,9 tikimybės slenkstį. Vienodomis spalvomis yra pažymėti junginiai ir su jais susije kristalografiniai parametrai. ”Fm3m” simetrijos grupė yra priskirta dviems junginiams. Raudonai pažymėtas parametras, kurio konvejeris neišgavo. Teksto šaltinis: [UMC+13]

Taip pat yra pastebima, kad parametrų išgavimo komponentas tam tikruose kontekstuose geba susieti junginio nuorodą su junginio pavadinimu. Nors nuorodų siejimas įvyksta ne visada tiksliai, šis reiškinys nurodo, kad šią užduotį galima išspręsti toliau plečiant ir tikslinant duomenų rinkinį.

> Influence of Pseudohalide Ions on the Molecular Structure and Magnetic Properties of the Manganese(II)-Bipyrimidine-Pseudohalide System. The compounds [Mn2(bpm)3(NCS)4] (1), [Mn(bpm)(NCO)2] (2), and [Mn2(bpm)(N3)4] (3) (where bpm= 2,2'-bipyrimidine) have been synthesized and characterized. **Compound 1** crystallizes in the triclinic P1 space group, with a = 9.122(1) Å, b = 9.229(1) Å, c = 11.710(2) Å, α = 74.89(1)°, β = 80.30(1)°, γ = 61.04(1)°, V = 831.7(2) Å3, Z = 2, R(Fo) = 0.029, and wR(Fo2) = 0.078. **Compound 2** crystallizes in the monoclinic C2/c space group, with a = 7.309(1) Å, b = 14.498(2) Å, c = 10.740(4) Å, β = 99.93(4)°, V = 1121.0(5) Å3, Z = 4, R = 0.025, and wR(Fo2) = 0.060. In the case of **compound 3**, the crystal parameters of the monoclinic C2/m space group are a = 6.206(2) Å, b = 15.084(2) Å, c = 8.909(1) Å, β = 95.75(2)°, V = 829.8(3) Å3, Z = 4, R = 0.029, and wR(Fo2)= 0.069.

18 pav. Parametrų išgavimas taikant SciBERT + CCD ir BioBERT + SQUAD2 + CQAD konvejerį ir 0,9 tikimybės slenkstį. Vienodomis spalvomis yra pažymėti junginiai ir su jais susije kristalografiniai parametrai. Raudonai pažymėtas parametras, kurio konvejeris neišgavo. Teksto šaltinis: [CUL+97]

7.1.2.6. Konvejerio lankstumas

Įprastai kristalografiniai tyrimai fiksuoja didesnę parametrų įvairovę, nei pateikta CQAD duomenų rinkinyje. Narvelio tūris, Z, lydymosi temperatūra, parametrai prie skirtingų sąlygų ir kitos savybės nebuvo įtrauktos į pirminį CQAD duomenų rinkinį. Išvardintų parametrų pildymas reikalautų kitokių automatinių žymėjimo metodų, nei buvo pritaikyti pasirinktiems parametrams. Pavyzdžiui, Z arba lydymosi temperatūros reikšmės iš 1-3 skaitmenų tekste gali būti nepakankamai unikalios, todėl tokių reikšmių žymėjimas 7.1.1 skyriuje aprašytu metodu sukurtų aibę klaidingai teigiamų pavyzdžių duomenų rinkinyje. Nors ir apmokymo duomenų rinkinys netalpina tam tikrų parametrų pavyzdžių, visgi klausimų-atsakymų komponentų lankstumas teikia modelio pritaikymo srities plėtimo galimybes. Pritaikymo srities galimybės buvo patikrintos naudojant nedidelį duomenų rinkinį talpinantį 58 Z ir narvelio tūrio parametrus. SciBERT + CCD ir BioBERT + SQUAD2 + CQAD ir BioBERT + CCD ir BioBERT + SQUAD2 + CQAD konvejeriai atliko užduotį atitinkamai 0,87 ir 0,72 tikslumu (atkūrimo rodikliai 0,12 ir 0,14).

Konvejeris	Viso spėjimų	Neteisingi	Teisingi	Viso parametrų	Atkūrimas	Tikslumas	F1
BioBERT(NER) BioBERT(QA)	8	**1**	7	58	12,1	**87,5**	21,2
SciBERT(NER) BioBERT(QA)	11	3	**8**	58	**13,8**	72,7	**23,2**

6 lentelė. Bandymų rezultatai

7.1.2.7. Konvejerio apribojimai

Darbe aprašyti kristalografinių parametro išgavimo konvejeriai geba dirbti su įvairiais kristalografiniais tekstais, skirtingomis junginių ir parametrų pateikimo būdais. Pasirinkta konvejerio struktūra tiesiogiai priklauso nuo duomenų rinkinių kokybės bei įvairovės. Bandymai sufleruoja, kad plečiant ir įvairinant CCD ir CQAD duomenų rinkinius konvejeris gebės tiksliau nuskaityti įprastus ir retos struktūros kontekstus. Taip pat konvejerio struktūra neriboja papildomų parametrų nuskaitymo.

Visgi, analizuojant laisva forma aprašytus kristalografinius tyrimus buvo identifikuoti keli pasirinkto konvejerio struktūros apribojimai:

- Apribojimai siejami su transformerių modelių architektūra ir realizacija. Kadangi modelis palaiko ribota įvestį (512 morfemų BERT atveju), šio kiekio gali nepakakti viso parametro konteksto nuskaitymui. Pavyzdžiui, paminėtas junginys teksto pradžioje ir jo parametras ilgos pastraipos pabaigoje nebus susieti. Nekeičiant architektūros, šią problemą galima būtų spręsti apdorojant tekstą slenkančio lango principu ir saugojant papildomą kontekstinę informaciją tolimesniems teksto apdorojimo žingsniams.

- Taip pat pateikta konvejerio struktūras nėra pritaikyta junginių parametrų nuskaitymui prie skirtingų sąlygų. Dalis kristalografinių tyrimų yra atliekami prie skirtingų temperatūrų ir spaudimo, todėl vienam junginiui kristalografinis parametras gali kisti. Šią problemą galima

spręsti taip pat klausimų-atsakymų modelio pagalba, tačiau tokiu atveju prireiktų papildomo pirminio teksto apdorojimo įvairioms sąlygoms identifikuoti ir surinktos informacijos pagalba galėtų būti formuojami sąlygas tikslinantys klausimai.

- Junginių pateikimo formos neapsiriboja konvenciniais IUPAC, SMILES, InChI ar molekulinėmis formulėmis. Kristalografiniuose tekstuose autoriai plačiai naudoja sutrumpinimus, abreviatūras, ”makro komandas” su nuorodomis. Pavyzdžiui, sakinyje ”*Cyanide-bridged bimetallic assemblies [Ni(pn)$_2$]$_2$[Fe(CN)$_6$]X·catnH_2O (X = ClO_4- and n = 2 (1); X = BF_4- and n = 2 (2); X = PF_6- and n = 2 (3)) ... have been prepared (pn = 1,2-propanediamine,* ...” yra minimi 3 cheminiai junginiai: *[Ni(1,2-propanediamine)$_2$]$_2$[Fe(CN)$_6$]ClO_4-·cat$_2H_2O$ (1)*, *[Ni(1,2-propanediamine)$_2$]$_2$[Fe(CN)$_6$]BF_4-·cat$_2H_2O$ (2)* ir *[Ni(1,2-propanediamine)$_2$]$_2$[Fe(CN)$_6$]PF_6-·cat$_2H_2O$ (3).* Automatizuoti tokio tipo junginių konvertavimą galima reguliariųjų išraiškų pagalba prieš atliekant likusius parametrų išgavimo žingsnius. Šis funkcionalumas esamame sprendime nėra įgyvendintas.

Rezultatai ir Išvados

Darbo eigoje buvo pateiktas kristalografinės informacijos konvejerio sprendimas sudarytas iš dviejų komponentų. Atskirų konvejerio komponentų tikslumas buvo tiriamas taikant skirtingus BERT architektūros kalbos modelius. Buvo identifikuota kalbos modelio atitikimo tiksliniam tekstui svarba. Taip pat išaiškinta teigiama bendrinių duomenų rinkinių įtaka klausimų-atsakymų modelio rezultatams. Darbo metu buvo sukurtas unikalus kristalografinių parametrų klausimų-atsakymų duomenų rinkinys, kuris talpina 5003 automatiškai sugeneruotus ir rankiniu būdu suformuotus įrašus. Taip pat buvo suformuotas išsamus viso konvejerio vertinimo duomenų rinkinys, kuris talpina 14 tikslinių tekstų ir 267 kristalografinius junginių parametrus. Pateikti kristalografinių duomenų išgavimo konvejeriai apdoroja vertinimo rinkinį panašiu tikslumu, tačiau aukščiausią tikslumo rodiklį parodė SciBERT(NER) + BioBERT(QA) konvejeris — 91,9 (63,7 atkūrimo rodiklis).

Taigi, konvejerio struktūra, sudaryta iš junginių identifikavimo ir junginio-parametro siejimo klausimų-atsakymų modelių parodė patenkinamus rezultatus ir tam tikrą lankstumą plečiant modelio pritaikymo sritį. Darbe buvo apibrėžta nauja kalbos apdorojimo užduotis, todėl gauto sprendimo tikslumą lyginti su panašioms užduotims skirtais sprendimais nėra visiškai teisinga. Gautas kristalografinių duomenų išgavimo konvejeris parodė aukštus tikslumo rezultatus, kurie suteikia pagrindą tikėtis, jog tolimesnis konvejerio tobulinimas leis automatiškai atlikti parametrų išgavimą ir jų talpinimą duomenų bazėje. Taip pat darbe buvo aprašytas automatinis duomenų rinkinio anotavimo metodas, kurio pagalba galima toliau plėsti klausimų-atsakymų modelio pritaikymo sritį. Pateiktas informacijos išgavimo konvejeris sprendžia dalį sudėtingų uždavinių (pvz. netiesiogiai įvardinto junginio atpažinimas), tačiau nekonvenciškai pateiktų junginių nuskaitymas ir parametrų-sąlygų sąryšiu atpažinimo uždaviniai lieka neišspręsti.

Literatūra

[ABB+19] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter ir Roland Vollgraf. FLAIR: an easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p.p. 54–59, Minneapolis, Minnesota. Association for Computational Linguistics, 2019-06. DOI: 10.18653/v1/N19-4010. URL: https://www.aclweb.org/anthology/N19-4010.

[AZ12] Charu C Aggarwal ir ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[BLC19] Iz Beltagy, Kyle Lo ir Arman Cohan. SciBERT: a pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p.p. 3615–3620, Hong Kong, China. Association for Computational Linguistics, 2019-11. DOI: 10.18653/v1/D19-1371. URL: https://www.aclweb.org/anthology/D19-1371.

[CCG+07] Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall ir kt. Learning alignments and leveraging natural logic. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p.p. 165–170, 2007.

[CMG05] T. E. Crumpton, J. F. W. Mosselmans ir C. Greaves. Structure and oxide ion conductivity in bi28re2o49, a new bismuth rhenium oxide containing tetrahedral and octahedral re(vii). *Journal of Materials Chemistry*, 15(1):164, 2005. DOI: 10.1039/b412108m.

[CUL+97] Roberto Cortés, M. Karmele Urtiaga, Luis Lezama, J. Luis Pizarro, M. Isabel Arriortua ir Teófilo Rojo. Influence of pseudohalide ions on the molecular structure and magnetic properties of the manganese(ii)−bipyrimidine−pseudohalide system. *Inorganic Chemistry*, 36(22):5016–5021, 1997. DOI: 10.1021/ic960969h. eprint: https://doi.org/10.1021/ic960969h. URL: https://doi.org/10.1021/ic960969h.

[DCL+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee ir Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DCL+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee ir Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding, 2019. arXiv: 1810.04805 [cs.CL].

[DDA+12] Nick Day, Jim Downing, Sam Adams, N. W. England ir Peter Murray-Rust. Crystaleye: automated aggregation, semantification and dissemination of the world's open crystallographic data. *Journal of Applied Crystallography*, 45:316–323, 2012. DOI: 10.1107/S0021889812006462.

[FGW+21] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura ir Eduard Hovy. A survey of data augmentation approaches for nlp, 2021. DOI: 10.48550/ARXIV.2105.03075. URL: https://arxiv.org/abs/2105.03075.

[For15] A. Dominic Fortes. Crystal structures of deuterated sodium molybdate dihydrate and sodium tungstate dihydrate from time-of-flight neutron powder diffraction. *Acta Crystallographica Section E*, 71(7):799–806, 2015-07. URL: https://doi.org/10.1107/S2056989015011354.

[GCT+23] Luke P. J. Gilligan, Matteo Cobelli, Valentin Taufour ir Stefano Sanvito. A rule-free workflow for the automated generation of databases from scientific literature, 2023. arXiv: 2301.11689 [cond-mat.mtrl-sci].

[GDM+23] Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner ir kt. Crystallography open database, 2023. URL: http://www.crystallography.net/cod/.

[HAB91] S. R. Hall, F. H. Allen ir I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A Foundations of Crystallography*, 47(6):655–685, 1991-11. DOI: 10.1107/s010876739101067x.

[HMS+13] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi ir Igor Pletnev. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1), 2013-01. DOI: 10.1186/1758-2946-5-7.

[JCP+] Ji-un Jeon, Do-Heon Choi, So-Young Park ir Yeo-Chan Yoon. Deep learning-based end-to-end dependency parsing without pos tagging.

[KHO98] Mitsuhiro Kataoka, Akio Hasebe ir Etsuo Otsuki. Crystal structure and magnetic anisotropy of sm2(fe,v)17+d. *Journal of the Magnetics Society of Japan*, 22(97):328–330, 1998. DOI: 10.3379/jmsjmag.22.S1_328.

[KKF+08] Corinna Kolárik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius ir Juliane Fluck. Chemical names: terminological resources and corpora annotation. *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, tom. 36, 2008.

[KKK+19] Yukari Katsura, Masaya Kumagai, Takushi Kodani, Mitsunori Kaneshige ir kt. Data-driven analysis of electron relaxation times in pbte-type thermoelectric materials. *Science and Technology of Advanced Materials*, 20(1):511–520, 2019. DOI: 10.1080/14686996.2019.1603885. eprint: https://doi.org/10.1080/14686996.2019.1603885. URL: https://doi.org/10.1080/14686996.2019.1603885.

[KOT+03] J.-D. Kim, T. Ohta, Y. Tateisi ir J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):i180–i182, 2003-07. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg1023. eprint: https://academic.oup.com/bioinformatics/article-pdf/19/suppl_1/i180/614820/btg1023.pdf. URL: https://doi.org/10.1093/bioinformatics/btg1023.

[KPM+13] Florian Kraus, Monalisa Panda, Thomas Müller ir Barbara Albert. Closo-hydroborates from liquid ammonia: synthesis and crystal structures of [li(nh3)4]2[b12h12]·2nh3, rb2[b12h12]·8nh3, cs2[b12h12]·6nh3 and rb2[b10h10]·5nh3. *Inorganic Chemistry*, 52(8):4692–4699, 2013. DOI: 10.1021/ic4002972. eprint: https://doi.org/10.1021/ic4002972. URL: https://doi.org/10.1021/ic4002972. PMID: 23537308.

[KRL+15] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez ir kt. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(S1), 2015-01. DOI: 10.1186/1758-2946-7-s1-s2.

[LCF+91] Wendy Lehnert, Claire Cardie, David Fisher, Ellen Riloff ir Robert Williams. University of Massachusetts: Description of the CIRCUS System as Used for MUC-3. Tech. atask., MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER ir INFORMATION SCIENCE, 1991.

[LYK+19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So ir Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019-09. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.

[LMS+93] Wendy Lehnert, Joe McCarthy, Stephen Soderland, Ellen Riloff, Claire Cardie, Jonathan Peterson ir Fangfang Feng. Umass/hughes: description of the circus system used for muc-51. *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.

[LSJ+16] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky ir kt. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[MH16] Xuezhe Ma ir Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016. arXiv: 1603.01354 [cs.LG].

[MN11] Ryan McDonald ir Joakim Nivre. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37:197–230, 2011-03. DOI: 10.1162/coli_a_00039.

[MR01] Peter Murray-Rust ir Henry S. Rzepa. Chemical markup, xml and the world-wide web. 2. information objects and the cmldom. *Journal of Chemical Information and Computer Sciences*, 41:1113–1123, 2001. DOI: 10.1021/ci000404a. eprint: http://pubs.acs.org/doi/pdf/10.1021/ci000404a. URL: http://pubs.acs.org/doi/abs/10.1021/ci000404a.

[NDG+16] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg ir kt. Universal dependencies v1: a multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p.p. 1659–1666, 2016.

[NKB+19] Mark Neumann, Daniel King, Iz Beltagy ir Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, p.p. 319–327, Florence, Italy. Association for Computational Linguistics, 2019-08. DOI: 10.18653/v1/W19-5034. eprint: arXiv:1902.07669. URL: https://www.aclweb.org/anthology/W19-5034.

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee ir Luke Zettlemoyer. Deep contextualized word representations, 2018. arXiv: 1802.05365 [cs.CL].

[Por80] Martin F Porter. An algorithm for suffix stripping. *Program*, 1980.

[RB21] Abigail Rai ir Samarjeet Borah. Study of various methods for tokenization. Jyotsna K. Mandal, Somnath Mukhopadhyay ir Alak Roy, redaktoriai, *Applications of Internet of Things*, p.p. 193–200, Singapore. Springer Singapore, 2021. ISBN: 978-981-15-6198-6.

[RJL18] Pranav Rajpurkar, Robin Jia ir Percy Liang. Know what you don't know: unanswerable questions for squad, 2018. DOI: 10.48550/ARXIV.1806.03822. URL: https://arxiv.org/abs/1806.03822.

[RZL+16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev ir Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. DOI: 10.48550/ARXIV.1606.05250. URL: https://arxiv.org/abs/1606.05250.

[SJP+18] Hye-Jeong Song, Byeong-Cheol Jo, Chan-Young Park, Jong-Dae Kim ir Yu-Seop Kim. Comparison of named entity recognition methodologies in biomedical documents. *Biomedical engineering online*, 17(2):1–14, 2018.

[SRJ+97] Richard P. Schutte, Steven J. Rettig, Ajey M. Joshi ir Brian R. James. Synthesis, structure, and reactivity of [rucl(pp)l]pf6 (pp = (pph3)2, ph2p(ch2)4pph2; l = p(py)3, pph(py)2, py = 2-pyridyl). the "missing" p,n,n'-coordination mode for 2-pyridylphosphines. *Inorganic Chemistry*, 36(25):5809–5817, 1997. DOI: 10.1021/

ic970835j. eprint: https://doi.org/10.1021/ic970835j. URL: https://doi.org/10.1021/ic970835j. PMID: 11670203.

[SS19] Sam Schwager ir John Solitario. Question and answering on squad 2.0: bert is all you need. *ArXiv e-prints of*, 2019.

[SSB+97] Yeung-gyo K. Shin, David J. Szalda, Bruce S. Brunschwig, Carol Creutz ir Norman Sutin. Electronic and molecular structures of pentaammineruthenium pyridine and benzonitrile complexes as a function of oxidation state. *Inorganic Chemistry*, 36(14):3190–3197, 1997. DOI: 10.1021/ic9700967. eprint: https://doi.org/10.1021/ic9700967. URL: https://doi.org/10.1021/ic9700967. PMID: 11669976.

[Str18] Milan Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p.p. 197–207, Brussels, Belgium. Association for Computational Linguistics, 2018-10. DOI: 10.18653/v1/K18-2020. URL: https://www.aclweb.org/anthology/K18-2020.

[Sun91] Beth M. Sundheim. Overview of the third message understanding evaluation and conference. *Proceedings of the 3rd Conference on Message Understanding*, MUC3 '91, p.p. 3–16, San Diego, California. Association for Computational Linguistics, 1991. ISBN: 1558602364. DOI: 10.3115/1071958.1071960. URL: https://doi.org/10.3115/1071958.1071960.

[ŠLC09] Petr Štěpnička, Martin Lamač ir Ivana Císařová. 1,2:5,6-Di-*O*-isopropylidene-α-D-3-glucofuranosyl (R_p)-2-(diphenylphosphino)ferrocene-1-carboxylate. *Acta Crystallographica Section E*, 65(10):m1252–m1253, 2009-10. URL: https://doi.org/10.1107/S1600536809038653.

[UMC+13] Christopher C. Underwood, Colin D. McMillen, Hongyu Chen, Jeffery N. Anker ir Joseph W. Kolis. Hydrothermal chemistry, structures, and luminescence studies of alkali hafnium fluorides. *Inorganic Chemistry*, 52(1):237–244, 2013. DOI: 10.1021/ic301760a. eprint: https://doi.org/10.1021/ic301760a. URL: https://doi.org/10.1021/ic301760a. PMID: 23245214.

[Wei88] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988-02. DOI: 10.1021/ci00057a005.

[WYK+20] Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu ir Jon Elsas. Learning to extract attribute value from product via question answering: a multi-task approach. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, p.p. 47–55, Virtual Event, CA, USA. Association for Computing Machinery, 2020. ISBN:

9781450379984. DOI: 10.1145/3394486.3403047. URL: https://doi.org/10.1145/3394486.3403047.

[WYP+11] Jian Wang, Min Yang, Ming-Yan Pan, Sheng-Qing Xia, Xu-Tang Tao, Hua He, Gregory Darone ir Svilen Bobev. Synthesis, crystal and electronic structures, and properties of the new pnictide semiconductors a2cdpn2 (a = ca, sr, ba, eu; pn = p, as). *Inorganic Chemistry*, 50(17):8020–8027, 2011. DOI: 10.1021/ic200286t. eprint: https://doi.org/10.1021/ic200286t. URL: https://doi.org/10.1021/ic200286t. PMID: 21786747.

[WSC+16a] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le ir kt. Google's neural machine translation system: bridging the gap between human and machine translation, 2016. DOI: 10.48550/ARXIV.1609.08144. URL: https://arxiv.org/abs/1609.08144.

[WSC+16b] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le ir kt. Google's neural machine translation system: bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.